

PRINSIP-PRINSIP PENGEMBANGAN KALIBRASI INSTRUMEN DALAM PENELITIAN

Roni Junaidi^{1*}, Risnita², M. Syahran Jailani³, Faisal Hakim Nasution⁴

¹²³⁴Universitas Islam Negeri Sulthan Thaha Saifuddin Jambi

*Corresponding Author: roni.junaidikamad@gmail.com

Abstrak

Penelitian merupakan suatu proses meneliti suatu fenomena/peristiwa secara sistematis yang ditujukan untuk menemukan dan/atau mengembangkan suatu pengetahuan yang benar. Sebagai suatu kegiatan sistematis penelitian harus dilakukan dengan metode tertentu yang dikenal dengan istilah metode penelitian, yakni suatu cara ilmiah yang dilakukan untuk mendapatkan data dengan tujuan dan kegunaan tertentu. Cara ilmiah tersebut harus didasari ciri-ciri keilmuan yaitu rasional, empiris, dan sistematis. Keberadaan instrumen penelitian merupakan bagian yang sangat integral dan termasuk dalam komponen metodologi penelitian untuk melaksanakan suatu kegiatan penelitian, karena instrumen penelitian merupakan alat yang digunakan untuk mengumpulkan, memeriksa, dan menyelidiki suatu masalah yang sedang diteliti. Suatu instrumen yang baik tentu harus memiliki validitas dan realibilitas yang baik. Untuk memperoleh instrumen yang baik tentu selain harus diuji cobakan, dihitung validitas dan realibilitasnya juga harus dibuat sesuai kaidah-kaidah penyusunan instrumen. Berkaitan dengan hal tersebut, pada pembahasan ini akan diuraikan berbagai hal terkait dengan instrumen penelitian yang pembahasannya diawali dengan pengertian instrumen penelitian, jenis, langkah-langkah penyusunan, dan kriteria instrumen yang baik.

Kata Kunci: Penelitian, Instrumen, Ilmiah

Abstract

Research is a systematic process of investigating a phenomenon/event aimed at discovering and/or developing true knowledge. As a systematic activity, research must be carried out with a specific method known as the research method, which is a scientific approach undertaken to obtain data for a specific purpose and utility. This scientific approach must be based on the characteristics of scholarship, namely rational, empirical, and systematic. The existence of research instruments is a highly integral part and is included in the methodology component of research to conduct research activities because research instruments are tools used to collect, examine, and investigate a problem under study. A good instrument must certainly have good validity and reliability. To obtain a good instrument, besides being tested and experimented, its validity and reliability must also be calculated according to the rules of instrument preparation. In connection with this, various aspects related to research instruments will be discussed in this discussion, starting with the understanding of research instruments, types, steps of preparation, and criteria for a good instrument.

Keywords: Research, Instrument, Scientific.

PENDAHULUAN

Instrumen penelitian adalah alat yang digunakan untuk mengumpulkan data penelitian, baik data yang kualitatif maupun kuantitatif. Data kualitatif dapat berupa gambar, kata, dan atau benda lainnya yang non angka, sedangkan data kuantitatif adalah data yang bersifat atau berbentuk angka. Dalam penelitian kualitatif instrumen utamanya adalah peneliti sehingga yang dimaksud dengan instrumen penelitian dalam kesempatan ini adalah instrumen penelitian kuantitatif.

Data kuantitatif itu sendiri dapat dikelompokkan menjadi dua, yaitu: data nominal dan data kontinum. Data dikatakan pada tingkat nominal atau berskala nominal apabila angka tersebut berfungsi untuk identifikasi, yaitu membedakan jenis subyek yang lainnya. Perbedaan angka hanya menunjukkan adanya obyek atau subyek yang terpisah dan tidak sama. Sementara itu data yang kontinum terdiri data yang berskala ordinal, interval, dan

rasio.

Data kuantitatif biasanya diperoleh melalui pengukuran, yaitu suatu proses pemberian angka pada subyek, obyek atau trait lainnya. Oleh karenanya instrumen penelitian dapat pula disebut dengan alat ukur, dan alat ukur ini dapat berupa tes dan nontes. Alat ukur dikatakan tes apabila memuat pertanyaan-pertanyaan yang jawabannya ada benar dan salah. Sebaliknya, alat ukur yang jawabannya tidak ada benar-salah dapat disebut dengan skala, angket, atau dapat pula disebut dengan inventori. Skala biasanya digunakan untuk mengukur konstruk atau konsep psikologis seperti: sikap, minat, motivasi, pendapat, dan *trait* lainnya, sedangkan angket digunakan untuk mengukur fakta, atau yang dianggap fakta seperti: pendidikan terakhir, jumlah anggota, penghasilan setiap bulan, dll. Sementara itu, inventori digunakan untuk mengungkapkan kepemilikan benda nyata, seperti: jumlah kursi, jumlah meja, dll.

Terdapat beberapa jenis desain penelitian dalam pendidikan. Creswell (2019) 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 organisas. 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 (2018), 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 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. Hal ini diperkuat juga oleh Azwar (2020) 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 (2018), paling tidak terdapat empat atribut afektif yang penting dalam pembelajaran, yaitu; sikap, minat, konsep diri, dan nilai. Sikap, menurut Ajzen (2017), 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.

METODE PENELITIAN

Untuk mendapatkan informasi yang akurat diperlukan metode yang tepat yang dikumpulkan dalam penelitian ini sebagai metode dokumenter. yaitu teknik perolehan data melalui pengumpulan dan analisis dokumen, seperti dokumen tertulis, gambar, karya dan elektronik. Dokumen yang diterima dianalisis, dibandingkan dan diintegrasikan (sintesis) menjadi suatu kajian yang sistematis, terpadu dan lengkap. Studi dokumenter tidak sekedar mengumpulkan dan merekam atau melaporkan dalam bentuk kutipan dari beberapa dokumen. Hasil penelitian yang dilaporkan merupakan hasil analisis terhadap dokumen-dokumen tersebut.

HASIL DAN PEMBAHASAN

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 merupakan 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 konstruk 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 kekonsistennya? Konsistensi hasil ukur inilah yang disebut sebagai reliabilitas (Crocker & Algina, 1986).

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 tester 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 alatukur 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.

Syarat utama instrumen yang baik adalah valid dan reliabel. Validitas suatu alat ukur adalah sejauhmana alat ukur itu mampu mengukur apa yang seharusnya diukur. Validitas pada umumnya bersifat tingkat bukan ada atau tidak ada sama sekali. Validitas suatu instrumen juga hanya dilihat dari tujuan tertentu; artinya suatu instrumen dikatakan valid untuk mengukur atribut A tidak harus valid untuk mengukur atribut B.

Dalam pengukuran terhadap atribut psikologis, validitas sebagaimana dijelaskan di atas sangat sulit dicapai. Hal ini dapat difahami karena pengukuran terhadap variabel psikologis dan sosial mengandung kesalahan yang lebih banyak daripada pengukuran variabel yang bersifat fisik. Oleh karena sulitnya menentukan validitas yang sebenarnya, maka yang dapat dilakukan adalah mengestimasi validitas instrumen dengan perhitungan tertentu.

Ada tiga tipe validitas, yakni: (1) validitas prediktif, (2) validitas isi, dan (3) validitas konstruk (Nunnally, 1978). Validitas prediktif atau ada juga yang menyebut dengan validitas kriteria terkait dicari manakala instrumen akan digunakan untuk mengestimasi beberapa bentuk tingkahlaku penting yang ada di luar dari hasil pengukuran instrumen itu sendiri. Atau, validitas prediktif diestimasi manakala instrumen dimaksudkan untuk berfungsi sebagai prediktor bagi performansi di waktu 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 obyek yang hendak diukur (aspek representasi) dan sejauh mana butir-butir itu mencerminkan ciri perilaku yang hendak diukur (aspek relevansi). Validitas isi suatu instrumen ditentukan dengan cara mencocokkan apakah butir-butir yang ada di instrumen itu sudah mewakili komponen-komponen yang akan diukur atau belum. Hal ini menunjukkan bahwa tingkat validitas isi suatu instrumen sedikit-banyak tergantung pada penilaian subyektif individual penilai. Hal ini diperkuat dengan adanya kenyataan bahwa estimasi validitas isi tidak melibatkan perhitungan statistik apapun melainkan hanya menggunakan analisis rasional.

Validitas konstruk adalah validitas yang menunjukkan sejauhmana instrumen mengungkap suatu trait atau konstruk teoritik yang hendak diukurnya. Pengujian validitas konstruk merupakan proses yang terus berlanjut sejalan dengan perkembangan konsep *trait* yang akan diukur. Perubahan dan perkembangan konsep seperti ini merupakan hal biasa dalam bidang psikologi karena variabel itu pada dasarnya merupakan konsep hipotetik yang tidak selalu mudah untuk dioperasionalkan. Konsep validitas konstruk sangat bermanfaat pada tes yang mengukur trait yang tidak memiliki kriteria eksternal. Untuk itu 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 konstruksi praktis mengenai hasil pengukuran pada kondisi tertentu, dan konstruksi inilah yang akan diuji. Apabila hasilnya sesuai dengan harapan maka instrumen itu dianggap memiliki validitas konstruk yang baik.

Prosedur validasi konstruk juga dapat ditempuh melalui teknik analisis faktor. Analisis faktor merupakan sekumpulan prosedur matematik yang cukup kompleks untuk menganalisis saling hubungan di antara variabel-variabel dan menjelaskan saling hubungan itu dalam bentuk kelompok variabel yang terbatas yang disebut faktor. Oleh karenanya validitas yang ditegakkan melalui prosedur analisis faktor disebut sebagai validitas faktorial. Syarat utama lainnya adalah instrumen itu harus reliabel. Sebenarnya reliabilitas itu mengacu pada konsistensi pengukuran, yaitu bagaimana skor tes atau hasil penilaian yang lain tetap (tidak berubah, sama) dari satu pengukuran ke pengukuran yang lain. Hasil-hasil penilaian hanya memberikan ukuran unjuk kerja terbatas yang diperoleh pada waktu tertentu. Kecuali kalau pengukuran dapat menunjukkan layak konsistensi atas kesempatan yang berbeda, penilai yang berbeda, atau sampel yang berbeda dan domain unjuk kerja yang sama. Hasil penilaian yang konsisten sempurna tidak mungkin dapat diperoleh. Banyak faktor yang mempengaruhi hasil penilaian.

Jika suatu penilaian dikenakan kepada kelompok yang sama dua kali secara berurutan, beberapa variasi skor dapat terjadi karena adanya fluktuasi pada memori sesat, perhatian, usaha kelelahan, ketegangan emosional, tebak-tebak dan sejenisnya. Sebaliknya jika dilaksanakan dalam waktu yang lama antara tes pertama dan tes kedua variasi skor kemungkinan disebabkan oleh pengaruh pengalaman belajar, perubahan kesehatan, lupa dan lain-lain. Variasi skor juga mungkin akan terjadi jika tes essay atau penilaian type unjuk kerja siswa lainnya yang dinilai oleh penilai yang berbeda.

Allen dan Yen (1979), mengemukakan tiga metode yang umum digunakan untuk menaksir koefisien reliabilitas yaitu: (1) metode tes ulang, (2) metode tes paralel, dan (3) metode konsistensi internal. Secara umum masing-masing dari ketiga metode tersebut akan menghasilkan taksiran koefisien reliabilitas (r_x), yang berbeda. Jadi yang dihasilkan hanyalah taksiran, karena nilai sebenarnya koefisien ini adalah tidak dapat diamati.

Sesuai dengan namanya, pada metode tes ulang pengambil tes yang sama mengikuti tes dua kali dengan menggunakan tes yang sama kemudian hasilnya dikorelasikan diperoleh taksiran reliabilitas. Metode tes ulang menghasilkan taksiran reliabilitas tes yang sangat beralasan, tetapi metode ini ternyata memiliki beberapa kelemahan. Pertama metode ini sangat potensial terpengaruh oleh *carry-over effect* antar tes, tes pertama sangat mungkin mempengaruhi hasil tes kedua. Kelemahan kedua berkenaan dengan waktu pelaksanaan tes. Interval waktu yang sangat pendek akan membuat *carry-over effect* dalam memori pengambil tes. Sedangkan interval waktu yang lama akan membawa pengaruh pada perubahan informasi.

Taksiran reliabilitas tes paralel adalah korelasi antara nilai amatan dua tes yang paralel. Dalam kenyataannya dua tes yang paralel hanyalah konsep teoritis, sangat sulit untuk membuktikan bahwa dua tes adalah paralel. Oleh karena itu sering digunakan bentuk tes alternatif sebagai pengganti tes paralel. Bentuk tes alternatif adalah setiap dua bentuk tes yang telah disusun dalam rangka untuk membuatnya paralel, dan keduanya mungkin memiliki rerata skor amatan, variansi, dan korelasi dengan pengukuran lain yang sama atau hampir sama. Korelasi antara skor amatan tes pertama dengan skor amatan tes alternatif r_{x2} adalah merupakan taksiran reliabilitas baik untuk tes pertama maupun tes alternatifnya.

Reliabilitas konsistensi internal ditaksir dengan satu kali pelaksanaan tes sehingga permasalahan yang menyertai metode tes ulang dapat dihilangkan. Metode untuk menaksir reliabilitas yang sangat luas telah diketahui dalam pendekatan ini adalah taksiran reliabilitas

belah dua. Pada pendekatan ini tes dibagi menjadi dua bagian, yang satu dan lainnya adalah dianggap sebagai tes alternatif, dan pembelahnya diatur sedemikian rupa sehingga keduanya merupakan tes paralel atau pada dasarnya □-ekuivalen.

Keuntungan utama penaksiran reliabilitas konsistensi internal adalah bahwa hanya diperlukan satu kali tes saja untuk menghitung taksiran reliabilitas. Namun demikian metodekonsistensi internal tidak cocok jika tes tidak dapat dibagi menjadi bagian-bagian yang paralel atau pada dasarnya □-ekuivalen atau jika tes tidak memiliki butir-butir independen yang dapat dipisahkan. Menurut Allen dan Yen (1979) ada tiga cara yang biasa digunakan untuk membelah tes menjadi dua bagian yaitu: (1) metode gasal-genap, butir-butir tes dikelompokkan berdasarkan butir-butir bernomor gasal dalam satu kelompok dan butir-butir tes bernomor genap ke dalam kelompok kedua, (2) metode belah dua sesuai dengan nomor urut.

Teknik untuk membelah tes menjadi dua dapat digeneralisasikan untuk membagi tes menjadi lebih dari dua komponen. Sebagai contoh metode gasal genap dapat dimodifikasi untuk membuat tiga komponen dari tes yang terdiri dari 9 butir dapat dikelompokkan menjadi, pertama butir nomor 1, 4 dan 7, kedua nomor 2, 5, dan 8, dan ketiga nomor 3, 6, dan 9. Berdasarkan asumsi tersebut maka pada dasarnya sebuah tes dapat dibagi menjadi N komponen di mana N maksimum adalah sebanyak jumlah butir dalam tes tersebut.

Langkah pertama yang harus dilakukan oleh peneliti adalah mengkaji secara teoritik tentang substansi yang akan diukur. Peneliti harus menentukan definisi konseptual kemudian definisi operasional. Selanjutnya definisi operasional ini dijabarkan menjadi indikator dan butir-butir. Menurut Tim Pusijsian (1997/1998, ada enam langkah untuk mengembangkan instrumen alat ukur, yaitu:

- 1) Menyusun spesifikasi alat ukur termasuk kisi-kisi dan indikator
- 2) Menulis pertanyaan
- 3) Menelaah pertanyaan
- 4) Melakukan ujicoba
- 5) Menganalisis butir instrumen
- 6) Merakit instrument dan memberi label

Spesifikasi alat ukur ini mencakup: tujuan pengukuran, kisi-kisi instrumen, skala pengukuran, dan panjang instrumen. Oleh karenanya dalam menentukan spesifikasi alat ukur berarti menentukan tujuan instrumen, mengembangkan kisi-kisi instrumen, menentukan skala pengukuran, dan menentukan panjang instrumen.

Di depan telah dikemukakan bahwa ada dua macam instrumen, yaitu instrumen untuk tes dan nontes. Oleh karenanya, perlu dibedakan antara kisi-kisi instrumen untuk tes dan kisi-kisi instrumen nontes. Secara rinci penyusunan kisi-kisi keduanya adalah sebagai berikut.

1. Kisi-kisi Instrumen/Tes

Setelah tujuan tes ditetapkan, kegiatan berikutnya adalah menyusun kisi-kisi tes. Kisi-kisi ini padadasarnya merupakan tabel matrik yang berisi spesifikasi soal yang akan ditulis. Kisi-kisi berisi tentang tujuan, standar kompetensi, kompetensi dasar, materi pokok, dan penilaian yang berisi bentuk dan jenis tagihan. Standar kompetensi dijabarkan menjadi kompetensi dasar, kompetensi dasar dipecah menjadi beberapa indikator, dan dari indikator inilah dibuat butir-butir instrumen.

Ada tiga langkah yang harus dipenuhi untuk menulis kisi-kisi, yaitu: 1) memilih standar kompetensi dasar, (2) memilih kompetensi dasar, (3) menulis indikator, dan (4) menentukan bentuk tes. Secara garis besar, ada dua bentuk tes yang banyak digunakan oleh guru, yaitu bentuk obyektif dan bentuk uraian atau nonobyektif. Sudah barang tentu, masing-masing bentuk tes memiliki kelebihan dan kekurangan.

2. Kisi-kisi Instrumen nontes

Penyusunan instrumen nontes didahului dengan penentuan definisi konseptual, kemudian dijabarkan lagi ke definisi operasional. Dari definisi operasional ini kemudian dijabarkan menjadi beberapa indikator yang selanjutnya dijabarkan menjadi butir-butir instrumen. Seperti yang telah dijelaskan di muka, instrumen nontes ini dibedakan menjadi

dua, yaitu skala, angket, dan inventori.

Skala digunakan untuk mengukur konstruk atau konsep psikologis seperti: sikap, minat, motivasi, pendapat, dan *trait* lainnya, sedangkan angket digunakan untuk mengukur fakta, atau yang dianggap fakta seperti: pendidikan terakhir, jumlah anggota, penghasilan setiap bulan, dll. Sementara itu, inventori digunakan untuk mengungkapkan kepemilikan benda nyata, seperti: jumlah kursi, jumlah meja, dll. Secara ringkas, hubungan antara tujuan, metode dan instrumen yang digunakan pada Tabel berikut.

Tabel 1 hubungan antara tujuan, metode dan instrumen yang digunakan pada penelitian

Tujuan untuk mengungkap:	Metode	Instrumen yg digunakan
- perilaku, kebiasaan, ketrampilan	observasi, wawancara mendalam	lembar observasi, lembar penilaian, catatan, peneliti sendiri
- potensi termasuk di dalamnya unjuk kerja	tes, perintah mengerjakan	soal tes, lembar perintah dilengkapi dg lembar observasi/ lembar penilaian
- afektif: motivasi, sikap, minat, kesukaan, dll	wawancara, survei	pedoman wawancara, skala
- data pribadi, data nyata	wawancara, survei	angket, inventori,
- data yang lalu, data sekunder	dokumentasi	daftar dokumen

Tabel di atas menjelaskan bahwa metode dan instrumen yang digunakan harus mengacu pada tujuan pengukuran. Hal ini penting agar tidak terjadi kesalahan pengukuran. Di muka telah dijelaskan pengertian dan jenis validitas dan reliabilitas instrumen. Secara ringkas cara memvalidasi dan mengestimasi reliabilitas instrumen dapat dilihat pada instrumen berikut.

Tabel 2 Teknis memvalidasi dan mengestimasi reliabilitas instrumen

Jenis Validitas	Cara Memvalidasi	Keterangan
Validitas isi: validitas kurikulum, validitas tampak	- menggunakan kisi-kisi - konsultasi keahlinya	-tanpa menggunakan teknik statistik
Validitas kriteria terkait atau validitas empirik: validitas prediktif, validitas konkuren	-mengkorelasikan dengan data di masa datang	Korelasi product moment
Validitas konstruk: validitas faktor	-mengkorelasikan skor butir dengan total	- analisis faktor - product moment - analisis butir
Jenis Reliabilitas	Prosedur	Teknik yang dipakai
Internal Consistency: 1. data ordinal 2. data nominal	1 dan 2, tes satu kali, kemudian dianalisis atau diestimasi reliabilitasnya 3 tes sekali, kemudian skor dibelah dua dan diestimasi	1. Koef. Alpha 2. KR 20, KR 21 3. Spearman Brown

Stabilitas	Tes dua kali dengan soal sama, kemudian hasilnya dikorelasikan.	Product moment dan korelasi intra kelas
Ekivalen	Beri tes dua kali dengan soal yang berbeda kemudian dikorelasikan	Product moment dan korelasi intra kelas

Tabel di atas menunjukkan bahwa untuk mengestimasi validitas dan reliabilitas instrumen diperlukan kerja yang sangat hati-hati, harus diupayakan agar proses dan estimasi ini dilaksanakan dengan sebaik-baiknya.

Untuk kasus ini atau untuk menyusun instrumen untuk mengukur kinerja SMK-SBI kali ini tidak perlu dituliskan kisi-kisi dan indikator karena sudah ada (WS 2). Selain itu, instrumen juga tidak perlu diuji coba dan analisis empirik karena memerlukan keahlian khusus dan memakan waktu tambahan. Jadi dalam kegiatan ini, yang harus dilakukan dalam penyusunan instrumen hanya menulis butir-butir instrumen dan menelaah butir. Setelah butir ditulis lalu ditelaah (diusahakan telaah dilakukan oleh orang lain atau bukan penulis butir).

Hal-hal yang harus diperhatikan adalah: (1) butir instrumen harus sesuai indikator, (2) butir ditulis secara singkat dan jelas, (3) pilihan jawaban yang berbentuk angka atau waktu, sebaiknya diurutkan, (4) dalam satu komponen, setiap butir diberi skor sama (skor sama tidak berarti pilihan jawabannya sama), dan (5) butir ditulis dengan menggunakan bahasa baku. Selain itu, untuk menarik responden agar mau merespon dengan baik maka instrumen sebaiknya: (1) dikemas dalam bentuk yang menarik, misal dalam bentuk buku yang agak kecil, (2) diusahakan jumlah butir untuk setiap jenis responden tidak terlalu banyak (maksimum 40 butir), dan (3) diusahakan butir pertanyaan dan jawaban pada halaman yang sama.

SIMPULAN DAN SARAN

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