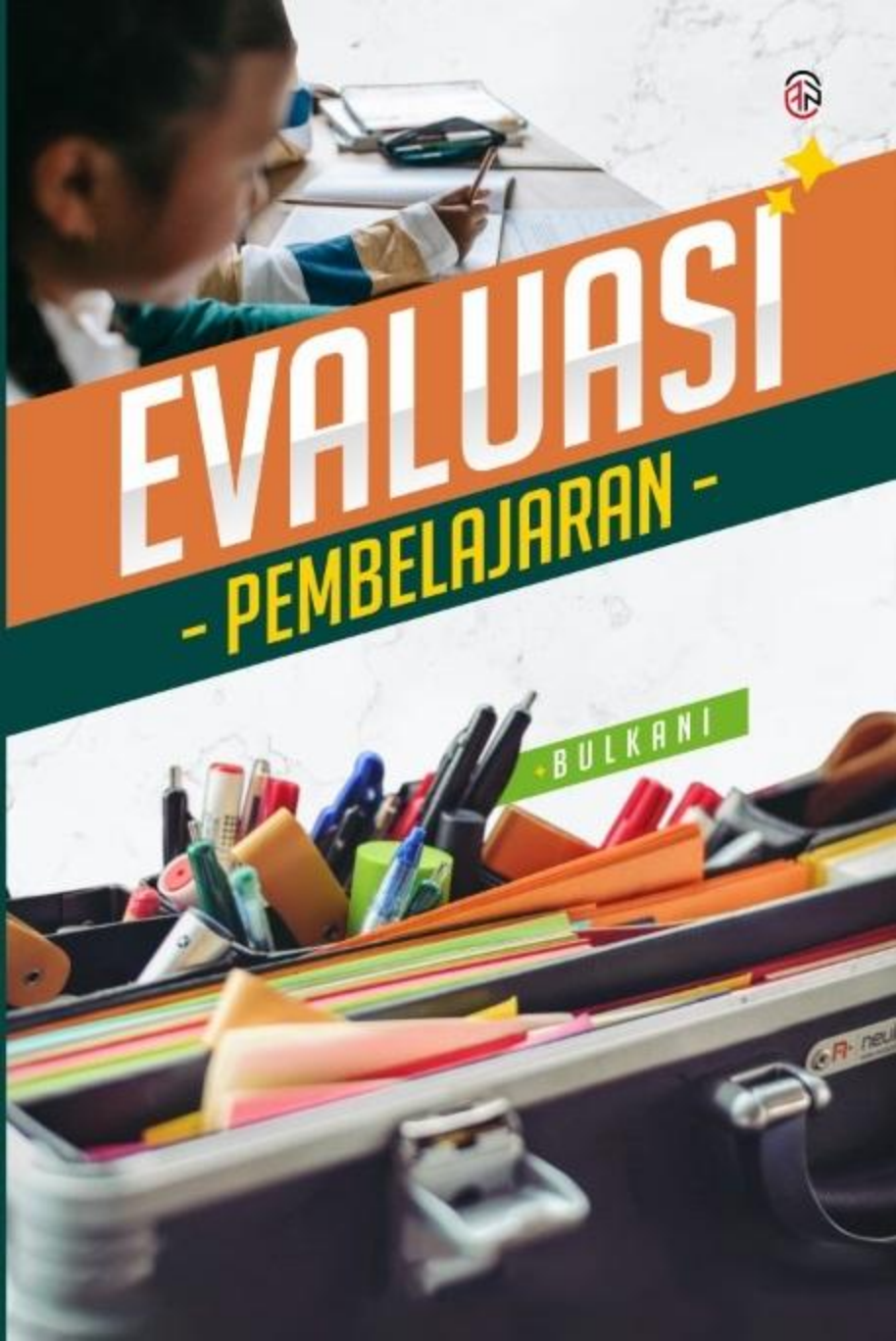
EVALUASI
- PEMBELAJARAN -

BULKANI



BULKANI

EVALUASI
PEMBELAJARAN

AKADEMIA
PUSTAKA

EVALUASI PEMBELAJARAN

Layouter: Muhamad Safi'i
Desain cover: Dicky M. Fauzi
Penyelaras akhir: Saiful Mustofa
x + 198 hlm: 14 x 21cm
Cetakan: Pertama, Oktober 2021
ISBN:

Anggota IKAPI

Diterbitkan oleh:
Akademia Pustaka
Perum. BMW Madani Kavling 16, Tulungagung
Telp: 081216178398
Email: redaksi.akademia.pustaka@gmail.com
Website: www.akademiapustaka.com

Kata Pengantar

Tulisan dan pembahasan tentang evaluasi pembelajaran selalu menarik untuk disajikan, baik secara teoretis maupun praktis. Paling tidak ada tiga penyebabnya. Pertama, ilmu evaluasi terus berkembang, sehingga menyebabkan terjadinya perubahan-perubahan teoretis maupun praktis dalam penggunaan evaluasi. Pandangan pandangan baru terhadap skor hasil pengukuran, khususnya hasil pembelajaran, juga mengalami perubahan. Skor hasil pengukuran yang selama ini dianggap sebagai sesuatu yang relatif stabil, cenderung berubah menjadi skor yang lebih dinamis, sehingga membutuhkan teknik pengukuran yang lebih akurat. Sebagai contoh, berkembangnya teori pengukuran, dari pendekatan teori skor klasik menjadi teori skor modern atau teori respon butir (*Item Respon Theory*), telah menyebabkan perubahan tentang instrumen dan analisis ujicobanya. Meskipun teori skor klasik masih sangat banyak digunakan saat ini dalam evaluasi, berkembangnya teori respon butir telah menyebabkan kehati-hatian dalam penggunaan instrumen dan analisisnya yang menggunakan teori skor klasik. Saat ini pengukuran hasil pembelajaran menggunakan teori skor klasik, lebih menekankan pada kehandalan instrumen yang tinggi, dengan pengujian yang lebih sering dan intensif dibandingkan dengan masa lalu.

Penyebab kedua adalah, berkembangnya masalah yang dihadapi oleh responden dan peserta evaluasi, seperti peserta didik. Dalam konteks pembelajaran, maka semakin kompleksnya masalah pembelajaran telah menyebabkan semakin kompleksnya perilaku hasil pembelajaran yang harus dievaluasi. Dispersi aspek perilaku hasil pembelajaran yang ada pada peserta didik juga menjadi semakin melebar rentangnya. Perkembangan ini membutuhkan studi dan

analisis baru yang mengakibatkan berkembangnya teori dan praktis evaluasi pembelajaran.

Ketiga, berubahnya keadaan di lingkungan di mana evaluasi dilaksanakan. Perubahan keadaan dianggap sebagai faktor yang berpengaruh terhadap hasil evaluasi. Hal ini membutuhkan teori dan praktik-praktik evaluasi yang lebih kontemporer.

Perubahan dalam teori dan praktik evaluasi pembelajaran, juga menimbulkan kebutuhan terus menerus terhadap buku yang membahas tentang evaluasi tersebut. Dalam buku ini dibahas tentang makna evaluasi, yang di dalamnya mengandung makna pengukuran dan penilaian. Pandangan masyarakat yang menyamakan antara pengukuran dan penilaian harus diluruskan. Demikian pula halnya tentang perbedaan skor dengan nilai. Pendekatan-pendekatan yang umumnya digunakan dalam mengukur dan menilai, serta bagaimana caranya merubah skor menjadi nilai, juga dibahas dalam buku ini. Selain itu, juga dibahas tentang pentingnya menyusun instrumen pengukuran yang handal, dan bagaimana cara memperolehnya. Pendekatan yang digunakan dalam pengukuran hasil pembelajaran pada buku ini, masih menggunakan teori skor klasik. Hal ini digunakan dengan alasan kepraktisan dan kegunaannya bagi para evaluator pemula.

Akhirnya, dengan berbagai kelebihan dan kekurangannya, buku ini diharapkan bermanfaat untuk menambah wawasan dan keterampilan baru kepada para evaluator dalam melaksanakan evaluasi pembelajarn di kelas. Semoga bermanfaat.

Palangka Raya, September 2021

Penulis,

Bulkani

Daftar Isi

Daftar Tabel

Daftar Gambar

BAB I
EVALUASI PEMBELAJARAN

A. Makna Evaluasi Pembelajaran

Evaluasi pada hakekatnya adalah aktivitas yang dilakukan secara sistematis oleh evaluator untuk melihat keberhasilan suatu kegiatan, yang dilakukan dengan cara membandingkan antara tujuan kegiatan dengan hasil yang telah dicapai oleh obyek tertentu. Melalui evaluasi, kita dapat mengetahui apakah suatu kegiatan yang dilaksanakan, telah atau belum mencapai tujuan yang diharapkan. Dengan kata lain, evaluasi adalah upaya membandingkan kesesuaian antara hasil yang dicapai dengan tujuan yang diharapkan (Mardapi, 2012). Stufflebeam & Sinkfield (2007), mendefinisikan evaluasi sebagai suatu proses menyediakan informasi yang dapat dijadikan sebagai bahan pertimbangan tentang ukuran-ukuran dari tujuan yang dicapai, desain, implementasi, dan dampak dari suatu kegiatan, yang dapat dijadikan sebagai bahan pengambilan keputusan tentang kegiatan tersebut. Sedangkan Widoyoko (2011), menyimpulkan bahwa evaluasi merupakan proses sistematis dan berkelanjutan untuk mengumpulkan, mendeskripsikan, menginterpretasi, dan menyajikan informasi tentang suatu program atau kegiatan.

Berdasarkan beberapa definisi di atas, maka evaluasi paling tidak mengandung beberapa makna sebagai berikut:

1. Evaluasi merupakan suatu proses yang sistematis dan berkelanjutan. Dalam konteks ini, evaluasi haruslah

merupakan proses yang terencana dan terukur langkah-langkahnya, sehingga langkah dan prosedurnya dapat ditelaah dan direplikasi oleh orang lain.

2. Evaluasi merupakan upaya membandingkan antara tujuan kegiatan dengan hasil yang dicapai.
3. Evaluasi memiliki obyek tertentu. Obyek evaluasi bisa berupa orang (misalkan peserta didik atau mahasiswa), program, atau benda.
4. Evaluasi membutuhkan evaluator. Evaluator adalah orang atau sekelompok orang yang memiliki keahlian khusus untuk melaksanakan evaluasi. Evaluator bisa berasal dari dalam (evaluator internal), maupun dari luar (evaluator eksternal).
5. Evaluasi menyediakan berbagai informasi tentang keberhasilan kegiatan, sehingga dapat dijadikan sebagai dasar kebijakan bagi keberlanjutan kegiatan tersebut.

Jika dikaitkan dengan definisi evaluasi di atas, maka evaluasi pembelajaran adalah suatu kegiatan yang dilakukan secara sistematis dan berlanjut untuk melihat hasil pembelajaran. Hasil pembelajaran dapat diartikan sebagai perubahan perilaku yang diharapkan akan terjadi pada peserta didik. Hasil belajar bukanlah tentang apa yang dilakukan guru, tetapi tentang apa yang diperoleh peserta didik diakhir program pembelajaran (Lestari & Setiawan, 2017).

Keberhasilan proses dan hasil pembelajaran tersebut dilihat dari perkembangan aspek-aspek hasil belajar, yang mencakup aspek kognitif, afektif, dan psikomotor. Evaluasi pembelajaran juga merupakan upaya untuk melihat tingkat pencapaian tujuan pembelajaran, sehingga dapat ditentukan beberapa alternatif tindak lanjut yang harus dilakukan terhadap peserta didik maupun sistem pembelajaran tersebut. Mansyur dkk (2009:8), mendefinisikan evaluasi pembelajaran sebagai proses pengumpulan informasi untuk mengetahui pencapaian belajar individu, kelas atau

kelompok. Sedangkan Widoyoko (2011:9) menyatakan bahwa evaluasi pembelajaran pada dasarnya adalah kegiatan yang dilakukan untuk melihat efektivitas program pembelajaran. Sedangkan Arifin (2012), berpandangan bahwa evaluasi pembelajaran merupakan kegiatan sangat penting dalam proses pembelajaran karena dapat memberikan informasi tentang kemajuan dan hambatan yang dialami peserta didik dalam belajar.

Dari beberapa uraian di atas, maka evaluasi pembelajaran mengandung makna sebagai berikut:

1. Evaluasi pembelajaran harus dilaksanakan secara sistematis, terencana, dan berkelanjutan.
2. Evaluasi pembelajaran bertujuan untuk melihat keberhasilan proses pembelajaran, yang dilakukan dengan cara membandingkan antara hasil pembelajaran dengan tujuan pembelajaran.
3. Adanya kriteria tertentu yang dijadikan dasar maupun patokan untuk menentukan keberhasilan obyek yang dievaluasi. Kriteria tersebut dapat bersifat kuantitatif maupun kualitatif.
4. Obyek evaluasi pembelajaran adalah peserta didik.
5. Evaluator dalam evaluasi pembelajaran adalah guru atau tenaga kependidikan lain yang ditunjuk.
6. Evaluasi pembelajaran menyediakan informasi tentang keberhasilan belajar peserta didik pada ranah kognitif, afektif, dan psikomotor. Informasi tersebut juga dapat digunakan untuk mendiagnosa kesulitan belajar yang dihadapi, menempatkan peserta didik pada posisi tertentu, dan melihat serta mendorong kemajuan belajar peserta didik.

B. Fungsi Evaluasi Pembelajaran

Sesuai dengan definisinya, maka evaluasi pembelajaran memiliki beberapa fungsi, yakni:

1. Fungsi penilaian

   Evaluasi pembelajaran berfungsi untuk menilai capaian hasil pembelajaran. Dari hasil evaluasi terhadap program pembelajaran, kita dapat menilai tercapai tidaknya tujuan program pembelajaran yang dilaksanakan. Dengan melakukan evaluasi pembelajaran, kita dapat menilai efektivitas program pembelajaran yang telah dilaksanakan. Efektivitas itu dapat dilihat dari besarnya perubahan tingkah laku peserta didik sebagaimana telah ditetapkan dalam tujuan pembelajaran.

   Meskipun demikian, fungsi evaluasi yang hanya berorientasi pada penilaian capaian tujuan pembelajaran, telah banyak mendapat kritik. Hal ini didasari pandangan bahwa program pembelajaran merupakan proses yang kompleks dan melibatkan banyak aspek, sehingga penilaian keberhasilannya tidak semata-mata dilihat dari pencapaian tujuan, tetapi juga harus melihat aspek-aspek lain seperti bagaimana input, sumberdaya, dan proses pembelajaran tersebut. Banyak fihak kemudian menganjurkan model evaluasi pembelajaran yang bebas dari tujuan (*goal free evaluation*). Mansyur dkk (2009:9), menyatakan bahwa program pembelajaran tidak semata-mata menghasilkan capaian tujuan belajar, tetapi juga bisa memperoleh hasil lain seperti tumbuhnya kepercayaan, keyakinan diri, kemandirian, tanggungjawab, dan aspek-aspek psikologis lainnya. Kirkpatrick (1998), menyatakan bahwa evaluasi pembelajaran yang baik haruslah melibatkan 3 komponen yang dievaluasi, yakni komponen pengetahuan yang dipelajari, keterampilan yang dihasilkan dan dikembangkan, serta sikap yang perlu diubah.

Dengan demikian, evaluasi pembelajaran akan berfungsi optimal jika evaluasi itu mencakup aspek-aspek psikologis secara komprehensif, yang dalam taksonomi Bloom dinamakan dengan aspek kognitif, afektif, dan psikomotor.

2. Fungsi diagnostik

Evaluasi pembelajaran juga berfungsi sebagai cara atau teknik untuk mendiagnosa kesulitan dan perkembangan belajar peserta didik. Dari sisi kesulitan belajar, hasil evaluasi pembelajaran dapat digunakan untuk mengetahui jenis dan peringkat kesulitan belajar peserta didik. Tingkat kesulitan belajar yang dialami peserta didik juga menggambarkan sejauhmana peserta didik tersebut telah mengalami perkembangan belajar. Widoyoko (2011; 34) menyatakan bahwa dari evaluasi, guru akan mengetahui kelebihan dan kelemahan serta kesulitan yang dialami peserta didik. Jika seorang peserta didik mengerjakan seperangkat soal evaluasi, maka pola jawabannya dapat menggambarkan tingkat pemahaman sekaligus tingkat kesulitan belajar yang dialaminya. Misalkan dalam evaluasi pembelajaran mata kuliah Matematika, seorang mahasiswa melakukan kesalahan dalam perhitungan, maka kesalahan tersebut menggambarkan kesulitan-kesulitan belajar mahasiswa tersebut dalam operasi hitung. Contoh lainnya adalah, kesalahan peserta didik kelas II sekolah dasar dalam memahami pertanyaan soal pada mata pelajaran Bahasa Indonesia, bisa menggambarkan adanya kesulitan peserta didik tersebut dalam mengenali huruf dan membaca.

3. Fungsi penempatan

Evaluasi pembelajaran juga dapat berfungsi untuk menempatkan peserta didik pada posisi tertentu, baik posisinya jika dibandingkan dengan peserta didik lain maupun terhadap standar tertentu. Dari hasil evaluasi,

guru dan pengambil kebijakan pendidikan dapat menempatkan seorang peserta didik pada kelas atau kelompok tertentu, menyatakan peserta didik belum atau telah tuntas belajarnya, menentukan seorang peserta didik telah layak diberi sertifiikat kelulusan, dan sebagainya. Widoyoko (2011; 34), menyatakan bahwa evaluasi pembelajaran dapat berfungsi sebagai dasar untuk menempatkan seorang peserta didik dalam kelompoknya, karena pembelajaran yang efektif adalah dengan memperhatikan homogenitas karakteristik individu dalam kelompok belajarnya.

Dalam konteks ini, evaluasi pembelajaran juga berfungsi selektif, artinya hasil evaluasi pembelajaran dapat digunakan oleh guru, sekolah dan pengambil kebijakan untuk menyeleksi peserta didik (Asrul dkk, 2015). Hasil seleksi tersebut digunakan sebagai dasar untuk menempatkan seorang peserta didik pada posisi kelulusan tertentu, pada kelas tertentu, melanjutkan atau berhenti pada titik tertentu karena tidak lulus seleksi.

4. Fungsi motivasi

Evaluasi pembelajaran juga dapat berfungsi sebagai alat memberi motivasi belajar. Hal ini didasari pada kenyataan bahwa peserta didik cenderung akan meningkatkan kesiapan belajarnya ketika akan dilakukan evaluasi belajar. Adanya keinginan untuk memperoleh nilai yang baik, cenderung menyebabkan peserta didik lebih serius untuk belajar.

Selain itu, peserta didik yang memperoleh nilai baik dari hasil evaluasi, akan memiliki motivasi semakin tinggi untuk mempertahankan dan meningkatkan nilai hasil belajarnya, sementara peserta didik yang memperoleh nilai kurang baik seharusnya termotivasi untuk mencapai nilai yang lebih baik pada evaluasi berikutnya.

C. Manfaat Evaluasi Pembelajaran

1. Bagi peserta didik

Tujuan pembelajaran bagi peserta didik adalah tercapainya perubahan tingkah laku kognitif, afektif, maupun psikomotorik. Untuk mengetahui tercapai tidaknya perubahan perilaku tersebut, maka dilakukan evaluasi pembelajaran, baik terhadap proses, hasil, maupun dampak pembelajaran. dengan demikian, manfaat utama evaluasi pembelajaran bagi peserta didik adalah untuk mengetahui kemajuan dan hasil belajarnya, yakni sampai di mana tujuan-tujuan belajar peserta didik telah tercapai. Widoyoko (2011; 36) menyatakan bahwa evaluasi hasil belajar akan membantu peserta didik mengetahui sejauhmana mereka berhasil mengikuti pelajaran yang diberikan guru. Evaluasi pembelajaran juga sering digunakan untuk mendiagnosa kesulitan-kesulitan belajar yang dialami peserta didik. Sementara Pauji dkk (2016), membuktikan dalam penelitiannya bahwa evaluasi pembelajaran bermanfaat untuk promosi, diagnosa kesulitan belajar, dan pengelompokkan peserta didik.

Dari hasil evaluasi pembelajaran, peserta didik dapat mengetahui perubahan perilaku seperti apa yang telah dicapainya, dan seperti apa seharusnya kemajuan dan hasil belajar tersebut dicapai. Evaluasi pembelajaran memberi manfaat bagi peserta didik sebagai cara untuk memetakan kemampuan mereka. Evaluasi pembelajaran juga bermanfaat untuk memotivasi peserta didik. Adanya keinginan untuk mencapai tujuan, posisi, dan peringkat tertentu dalam belajar, akan mendorong peserta didik untuk mempersiapkan diri dengan lebih baik saat dievaluasi. Leenknecht dkk (2020), menemukan bahwa evaluasi formatif yang diberikan pada peserta akan meningkatkan pengharapan mereka tentang pemuasan kebutuhan otonomi dan pencapaian kompetensi yang akhirnya akan meningkatkan motivasi belajarnya. Selegi

(2017) menemukan bahwa pelaksanaan evaluasi pembelajaran berpengaruh signifikans terhadap motivasi belajar mahasiswa. Penelitian Asdam (2007) menyimpulkan bahwa pemberian ulangan harian dalam frekuensi tertentu mampu secara signifikans meningkatkan motivasi belajar peserta didik.

2. Bagi guru

   Sebagai pendidik yang merencanakan dan melaksanakan pembelajaran, maka guru memiliki kepentingan sangat besar terhadap pelaksanaan evaluasi pembelajaran, karena evaluasi pembelajaran dapat memberikan gambaran tentang efektivitas proses pembelajaran yang diberikan guru. Sebagai pelaksana atau manajer pembelajaran, guru perlu mengetahui apakah pembelajaran yang dilaksanakannya telah merubah perilaku peserta didik sebagaimana diharapkan. Hal itu dapat diketahui dari hasil evaluasi pembelajaran. Hasil evaluasi pembelajaran dapat digunakan guru sebagai umpan balik untuk perbaikan pembelajaran (Pauji dkk, 2016). Widoyoko (2011; 38) menyatakan bahwa terdapat 3 manfaat evaluasi pembelajaran bagi guru, yakni (1). Mengetahui peserta didik yang mencapai kompetensi atau KKM yang diharapkan, (2). Mengetahui tingkat ketepatan pengalaman belajar atau materi pelajaram yang diberikan, (3). Mengetahui efektivitas strategi, media, dan teknik pembelajaran yang digunakan. Mardapi (2012, 14) menyatakan bahwa seharusnya evaluasi pembelajaran mampu memotivasi guru untuk melaksanakan pembelajaran dengan lebih baik.

3. Bagi sekolah

   Evaluasi pembelajaran juga memberi manfaat bagi sekolah. Pada hakekatnya, tercapainya tujuan pembelajaran oleh peserta didik, juga merupakan tujuan pendidikan bagi sekolah, yang disebut sebagai tujuan

institusional. Artinya, dengan mengetahui keberhasilan pembelajaran dan capaian tujuan belajar, maka sekolah juga mengetahui tercapai tidaknya tujuan institusionalnya sebagai lembaga penyelenggara program pendidikan. Hasil belajar peserta didik merupakan cermin dari kualitas dan keberhasilan tujuan sekolah (Widoyoko, 2011).

Dalam jangka panjang, kumpulan hasil evaluasi pembelajaran bermanfaat bagi sekolah sebagai bahan kebijakan dalam pengembangan dan orientasi jenis pendidikan. Menurut Mardapi (2012; 14), evaluasi pembelajaran idealnya mampu mendorong sekolah untuk mencapai kinerja dan kualitas pendidikan yang lebih baik. Asrul dkk (2015) berpendapat bahwa evaluasi pembelajaran bermafaat bagi sekolah untuk merencanakan program pendidikan dan menilai ketepatan kurikulum yang digunakan. Selain itu, kumpulan hasil evaluasi pembelajaran dalam jangka waktu tertentu dapat digunakan sekolah sebagai sarana promosi dan publikasi.

4. Bagi orangtua

Evaluasi pembelajaran sangat bermanfaat bagi orangtua untuk melihat keberhasilan pendidikan anak-anaknya. Salah satu manfaat evaluasi pembelajaran adalah tersusunnya laporan kemajuan belajar untuk orangtua peserta didik (Pauji dkk, 2016). Tujuan orangtua mempercayakan atau menitipkan pendidikan anak-anaknya kepada pihak sekolah adalah agar tercapai perkembangan belajar. Berkembang tidaknya perilaku belajar peserta didik diukur dari evaluasi pembelajaran. Dari hasil evaluasi pembelajaran, orangtua mengetahui apakah anak-anaknya telah memperoleh hasil belajar yang diharapkan. Hasil evaluasi pembelajaran di sekolah dapat digunakan oleh orangtua untuk membimbing kegiatan belajar anaknya di rumah. Bahkan dari hasil

evaluasi pembelajaran, orangtua dapat merencanakan arah pendidikan dan karir anak-anaknya di masa depan.

D. Jenis Evaluasi Pembelajaran

Secara umum, evaluasi pembelajaran dapat dibedakan menjadi evaluasi formatif dan evaluasi sumatif. Pembagian ini didasarkan pada waktu pelaksanaan, kebutuhan, dan cakupan materi evaluasi yang dilaksanakan.

1. Formatif

   Evaluasi formatif adalah evaluasi pembelajaran yang dilakukan untuk mengukur capaian sebagian materi pembelajaran yang telah diajarkan, baik untuk satu kali pertemuan, satu pokok bahasan, atau beberapa pokok bahasan.

   Termasuk jenis evaluasi formatif antara lain, evaluasi yang diberikan pada akhir pembelajaran atau ulangan harian. Demikian pula evaluasi yang dilakukan pada saat proses pembelajaran berlangsung. Evaluasi formatif diarahkan untuk memperbaiki bagian tertentu atau sebagian besar pembelajaran (Arifin, 2012). Evaluasi formatif juga dapat berfungsi sebagai penyedia data tentang kemajuan belajar dan hambatan belajar peserta didik, sehingga dapat dijadikan sebagai umpan balik oleh guru (Cyrs, 2010).

   Jika dikaitkan dengan aspek-aspek perilaku hasil belajar, maka evaluasi formatif juga dapat mencakup evaluasi terhadap aspek kognitif, afektif, dan psikomotorik. Kadangkala pengajar perlu melakukan pengamatan terhadap perubahan perilaku peserta didik, misalnya tentang bagaimana aktivitas, motivasi, minat dan sikap peserta didik saat pembelajaran. Evaluasi semacam ini merupakan bagian dari evaluasi formatif. Hasil evaluasi formatif kemudian dijadikan salah satu dasar pertimbangan dalam penentuan hasil evaluasi akhir.

2. Sumatif

Evaluasi sumatif adalah evaluasi pembelajaran yang dilaksanakan setelah satu paket pembelajaran telah diberikan. Evaluasi sumatif akan memberikan data tentang penguasaan peserta didik pada suatu materi pelajaran, menentukan posisi mereka terhadap peserta didik lain atau terhadap criteria tertentu (Cyrs, 2010). Evaluasi ini biasanya dilaksanakan di akhir program pembelajaran, misalnya pada akhir semester, saat kenaikan kelas, atau saat kelulusan. Ujian akhir semester dan ujian kelulusan, dapat digolongkan ke dalam jenis evaluasi sumatif. Dengan demikian evaluasi sumatif dihubungkan dengan penyimpulan mengenai perbaikan dari sistem kurikulum secara keseluruhan (Arifin, 2012).

Ditinjau dari substansi materinya, evaluasi sumatif mencakup materi pembelajaran yang lebih luas dan komprehensif dibandingkan evaluasi formatif. Evaluasi sumatif tidak hanya semata-mata mengumpulkan data dari pelaksanaan tes saat ujian semester semata-mata, tetapi juga mencakup pengumpulan data kemajuan belajar peserta didik dari hasil evaluasi formatif. Hasil evaluasi formatif biasanya juga digunakan sebagai komponen yang diperhitungan dalam evaluasi sumatif.

E. Prinsip-prinsip Evaluasi Pembelajaran

Evaluasi pembelajaran yang baik, dilaksanakan dengan prinsip-prinsip tertentu, yakni

1. Komprehensif

Evaluasi pembelajaran yang baik, haruslah merupakan evauluasi yang menyeluruh atau komprehensif, baik dari sisi aspek perilaku yang diukur, keterwakilan karakteristik obyek yang diukur, maupun substansi materi yang dievaluasi. Misalkan untuk mengevaluasi hasil belajar kita menggunakan seperangkat tes, maka butir-butir tes tersebut harus mampu mengukur

keseluruhan materi pelajaran yang telah diajarkan. Butir-butir tes tersebut juga harus mampu mengukur keseluruhan tujuan pembelajaran yang telah ditetapkan.

2. Kontinyu

   Evaluasi pembelajaran, harus memenuhi azas kesinambungan atau kontinyuitas. Hasil pembelajaran umumnya berbentuk aspek-aspek psikologis yang bersifat laten, sehingga perkembangannya sangat dinamis. Evaluasi yang kontinyu akan memberikan gambaran lebih akurat tentang perubahan perilaku yang telah dicapai peserta didik dalam pembelajaran.

3. Obyektif

   Evaluasi pembelajaran haruslah obyektif, adil, dan tidak pilih kasih. Evaluator harus memberikan penilaian apa adanya terhadap perilaku hasil belajar yang diperoleh peserta didik setelah mengikuti pembelajaran. Obyektivitas evaluator akan sangat membantu memberikan gambaran yang lebih nyata tentang hasil belajar, termasuk kemungkinan adanya hambatan-hambatan belajar, sehingga dapat ditentukan tindak lanjut dan perbaikan sistem pembelajaran.

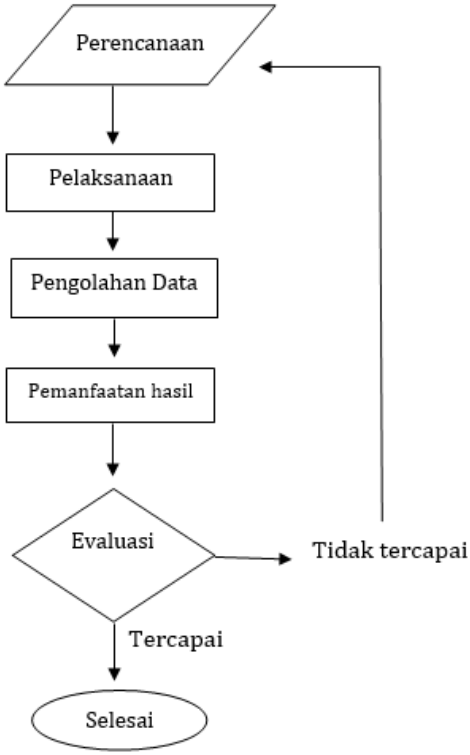
4. Akomodatif

   Evaluasi pembelajaran juga harus memenuhi prinsip akomodatif. Artinya, evaluasi tersebut harus dirancang untuk dapat mengakomodir pengukuran terhadap semua karakteristik peserta didik yang bervariasi. Adanya perbedaan variasi karakteristik peserta didik, menyebabkan perbedaan pula pada model dan pendekatan evaluasi pembelajaran yang akan digunakan.

F. Tahapan Evaluasi Pembelajaran

Berdasarkan pendekatan sistem, dan mengacu pada sistem manajemen, maka sistem evaluasi dapat dibagi 3 tahapan, yakni perencanaan (*planning),* pelaksanaan (*implementation),*

dan evaluasi (*evaluating).* Tahapan evaluasi pembelajaran diilustrasikan seperti diagram alur berikut:



Gambar 1.1. Diagram tahapan evaluasi pembelajaran

Diagram tersebut dapat dijelaskan sebagai berikut:

1. Perencanaan

   Setiap kegiatan harus dimulai dari perencanaan, demikian pula halnya dengan evaluasi. Perencanaan yang baik sebelum melaksanakan evaluasi, sangat berguna mengarahkan proses dan hasil evaluasi kearah yang benar, sehingga efektivitasnya dapat dijamin. Dengan

perencanaan yang baik, kegiatan evaluasi tersebut dapat dievaluasi sehingga ditemukan pada titik mana kelemahan dan kekurangan terjadi, yang sangat bergunan untuk penyempurnaan sistem evaluasi secara keseluruhan.

Beberapa hal pokok yang harus direncanakan dengan baik adalah:

a. Menentukan tujuan evaluasi. Evaluator harus menetapkan untuk apa evaluasi tersebut dilakukan. Langkah ini merupakan tahapan yang paling penting. Dalam tahapan ini, evaluator harus meninjau kembali keseluruhan tujuan kegiatan, sehingga dapat ditentukan tujuan mana yang akan dievaluasi. (Callahan & Logan, 2021).
b. Obyek yang akan dievaluasi. Obyek evaluasi dapat berupa program, kegiatan, peserta didik, peserta kegiatan, dan semacamnya.
c. Menentukan model evaluasi yang akan digunakan yang tergantung dari tujuan dan obyek yang dievaluasi. Jika menggunakan model evaluasi program, maka evaluator dapat memilih salah satu model evaluasi program, antara lain model CIPP (*Contex-Input-Proocess-Product*) dan CIPPO (*Contex-Input-Proocess-Product-Outcome*). Sedangkan jika evaluasi yang akan dilakukan adalah evaluasi pembelajaran denga obyek peserta didik, maka evaluator harus menentukan apakah evaluasi tersebut merupakan evaluasi formatif, sumatif, atau diagnostik.
d. Aspek-aspek apa yang akan diukur. Dalam konteks evaluasi pembelajaran, evaluator harus menentukan aspek dan sub aspek hasil belajar yang akan diukur.
e. Tingkat kedalaman dan keluasan aspek yang akan diukur. Untuk evaluasi dengan aspek-aspek hasil belajar yang mendalam, dibutuhkan teknik dan

instrumen yang berbeda dibandingkan dengan aspek yang lebih dangkal.

f. Jenis instrumen yang akan digunakan. dalam evaluasi, terdapat beberapa pilihan jenis dan bentuk instrumen yang dapat digunakan, tergantung pada tujuan evaluasi, obyek evaluasi beserta karakteristiknya, aspek yang akan diukur, dan beberapa faktor lainnya.

g. Waktu pelaksanaan evaluasi. Evaluator harus merencanakan waktu yang tepat untuk melaksanakan evaluasi sesuai dengan kebutuhan dan situasi yang dihadapi.

h. Personal yang terlibat dalam evaluasi. Kadangkala dalam pelaksanaan evaluasi yang cukup luas, evaluator perlu membentuk tim untuk merencanakan dan melaksanakan evaluasi tersebut. Untuk itu evaluator perlu menetapkan jumlah dan peran masing-masing anggota tim, sehingga setiap orang dapat berkontribusi dengan baik.

i. Biaya dan sumber biaya. Dalam konteks evaluasi yang luas dan melibatkan banyak obyek evaluasi yang banyak, evaluator kadangkala perlu merumuskan biaya yang dibutuhkan untuk melaksanakan evaluasi tersebut.

2. Pelaksanaan

Kualitas dan akurasi hasil evaluasi juga dipengaruhi oleh kelancaran pelaksnaan evaluasi. Itulah sebabnya evaluator harus merencanakan dengan matang pelaksanaan evaluasi. Faktor-faktor seperti karakteristik obyek yang akan dievaluasi, dan ketersediaan sarana prasarana, merupakan faktor yang harus dipertimbangkan. Prediksi terhadap situasi yang akan dihadapi ketika melaksanakan evaluasi, juga menentukam kualitas dan akurasi hasil evaluasi.

3. Pengolahan data

   Setelah data hasil evaluasi terkumpul, maka evaluator harus melakukan analisis terhadap data tersebut. Jenis analisis tersebut harus direncanakan dari awal. Analisis hasil evaluasi umumnya dapat dibedkan berdasarkan pendekatan yang digunakan, yakni analisis data kuantitatif dan analisis kualitatif. Evaluasi yang menghasilkan data berupa angka-angka atau skor, umumnya membutuhkan analisis kuantitatif berupa analisis statistika. Sedangkan evaluasi yang menghasilkan data kualitatif, juga membutuhkan analisis kualitatif. Pada kenyataannya di lapangan, evaluator sering menggunakan kedua pendekatan tersebut, sehingga hasil evaluasi lebih bermakna.

   Untuk evaluasi pembelajaran, pengolahan data juga tergantung dari tujuan dan pendekatan penilaian yang digunakan. Dalam hal ini, evaluator dapat menggunakan pendekatan PAN (Penilaian Acuan Normatif), dan dapat juga menggunakan pendekatan PAP (Penilaian Acuan Patokan). Pendekatan PAN digunakan jika evaluator bermaksud membandingkan dan menentukan posisi seorang peserta didik terhadap posisi peserta didik dalam kelas atau kelompoknya. Sedangkan pendekatan PAP digunakan jika evaluator bermaksud menentukan posisi seorang peserta didik terhadap batasan atau standar tertentu, misalnya dalam pengukuran kompetensi.

4. Pemanfaatan sesuai fungsi dan manfaat.

   Evaluator harus merencanakan pemanfaatan hasil evaluasi. Pada saat merencanakan evaluasi, evaluator harus menetapkan untuk apa hasil evaluasi ini akan dimanfaatkan. Hasil evaluasi dapat dimanfaatkan sesuai dengan fungsi evaluasi, yakni fungsi penilaian, fungsi

diagnostik, fungsi penempatan, dan fungsi motivasi. Evaluator dapat mengharapkan sebagian atau semua fungsi tersebut sebagai manfaat evaluasi.

Selain itu, evaluator juga harus merencanakan tentang siapa yang dapat memanfaatkan hasil evaluasi tersebut, apakah hasil evaluasi bermanfaat bagi peserta didik, bagi guru, bagi sekolah, dan bagi orangtua.

5. Evaluasi

Yang dimaksud dengan tahap evaluasi dalam hal ini adalah, melakukan penilaian terhadap keberhasilan dari tahapan evaluasi yang dilakukan oleh evaluator. Kadangkala dapat terjadi ketidaksesuaian antara tujuan evaluasi dengan hasil evaluasi yang dicapai. Dengan kata lain, tujuan evaluasi tidak dapat dicapai karena adanya berbagai hambatan. Sebagai contoh, ketika merencanakan dan melaksanakan evaluasi hasil pembelajaran, guru ingin mengetahui sebaran kompetensi peserta didik dalam rentang nilai yang merata, yakni sangat kompeten-cukup kompeten-kurang kompeten-tidak kompeten. Setelah dilakukan evaluasi, ternyata sebagian besar peserta didik termasuk dalam kategori kurang kompeten, sehingga guru akan membuat dugaan penyebabnya, antara lain karena sistem pembelajaran yang kurang efektif, atau bisa jadi instrumen dan proses evaluasinya yang kurang baik. Dengan demikian, guru selaku evaluator akan menilai kembali kualitas pembelajaran, dan mengevaluasi sistem evaluasi pembelajaran yang telah dilakukan. Dari hasil evaluasi terhadap sistem evaluasi pembelajaran tersebut, guru mungkin dapat mengidentifikasi beberapa penyebab atau kelemahannya.

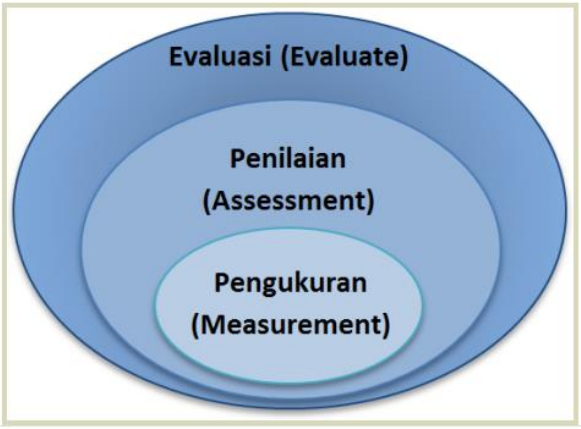
Jika yang dievaluasi adalah program pembelajarannya, maka guru dapat melakukan penilaian terhadap program secara keseluruhan, misalnya menggunakan model evaluasi CIPP. Jika tujuan program pembelajaran tidak

tercapai, maka guru selaku evaluator dapat menentukan titik kelemahan dari program pembelajaran tersebut, antara lain dari sisi sistem evaluasi yang digunakan, mungkin model evaluasi yang digunakan sebagai pendekatan kurang cocok terhadap situasi pembelajaran tersebut, dan sebagainya. Evaluasi semacam ini penting untuk perbaikan program pembelajaran dan sistem evaluasinya pada masa mendatang (Widoyoko, 2011).

Dari hasil evaluasi terhadap sistem evaluasi yang dilakukan evaluator, akan diperoleh 2 keputusan penting. Jika evaluasi tersebut dianggap telah mencapai tujuannya, maka sistem evaluasi tersebut dapat disimpan dan dibakukan untuk digunakan pada waktu yang akan datang pada situasi yang relevan. Jika ternyata evaluasi tersebut belum mencapai tujuan yang diharapkan, maka evaluator akan melakukan telaah terhadap sistem dan tahapan yang telah dilakukan, untuk menemukan hambatan dan kelemahannya sehingga tidak terjadi lagi pada evaluasi berikutnya.

BAB II
PENGUKURAN, PENILAIAN, DAN EVALUASI

Kata evaluasi tidak dapat dilepaskan maknanya dari pengukuran dan penilaian. Evaluasi selalu melibatkan pengukuran dan penilaian. Pengukuran dan penilaian merupakan tahapan dari proses evaluasi. Hubungan antara pengukuran, penilaian, dan evaluasi, dapat digambarkan melalui diagram berikut:



Gambar 2.1. Hubungan pengukuran, penilaian, dan evaluasi

Dari gambar di atas, dapat dijelaskan bahwa pengukuran merupakan tahap awal dari proses evaluasi. Hasil pengukuran kemudian dijadikan sebagai dasar untuk melakukan penilaian atau memberi nilai. Dari hasil penilaian, evaluator dapat

membuat keputusan akhir yang menggambarkan hasil evaluasi.

Perbedaan antara pengukuran dan penilaian dijelaskan sebagai berikut:

A. Pengukuran dan Skor

Asal kata pengukuran adalah kata "ukur" dan "mengukur". Pengukuran adalah mengukur suatu obyek, yakni membandingkan ukuran suatu obyek yang diukur, dengan menggunakan alat ukur standar yang telah disepakati/diketahui kehandalannya dan digunakan sebagai patokan. Misalkan kita melakukan pengukuran panjang sebuah meja, dengan cara membandingkan ukuran meja tersebut terhadap meteran. Meja adalah obyek yang diukur, sedangkan meteran merupakan ukuran standar tertentu yang digunakan sebagai patokan yang standar yang digunakan sebagai pembanding atau acuan. Contoh lain adalah ketika kita mengukur suhu tubuh seseorang menggunakan thermometer. Dalam hal ini tubuh merupakan obyek, suhu tubuh merupakan ukuran yang ingin diketahui, sedangkan thermometer merupakan alat ukur standar yang digunakan sebagai patokan. Pengukuran merupakan kegiatan untuk menentukan kuantitas pada sesuatu obyek (Arifin, 2012). Pengukuran merupakan upaya melakukan kuantifikasi atau penetapan nilai terhadap sesuatu obyek menggunakan instrumen tertentu (Mardapi, 2012). Pengukuran merupakan suatu proses pemberian angka kepada suatu atribut atau karaketristik tertentu yang dimiliki oleh orang, hal, atau obyek tertentu menurut aturan atau formulasi tertentu (Mansyur, dkk, 2009).

Dengan demikian, ada 3 hal yang selalu terlibat dalam proses pengukuran atau kegiatan mengukur, yakni adanya obyek yang diukur, adanya ukuran obyek yang ingin diketahui, adanya alat ukur standar yang digunakan sebagai pembanding atau sebagai patokan.

Dalam pengukuran obyek berupa benda fisik, umumnya kita memiliki ukuran dan alat ukur standar yang dijadikan sebagai patokan yang telah disepakati secara umum. Misalkan meteran merupakan alat ukur standar yang dijadikan patokan untuk mengukur panjang, thermometer merupakan alat ukur standar untuk mengukur suhu, timbangan merupakan alat ukur standar untuk mengukur berat, althometer merupakan alat ukur standar yang digunakan untuk mengukur ketinggian, dan sebagainya.

Hal berbeda akan kita temukan pada pengukuran aspek-aspek psikologis, seperti pengukuran hasil belajar peserta didik. Hasil belajar merupakan aspek psikologis yang tidak tampak secara fisik. Dalam konteks ini, obyek ukurnya bisa tampak secara fisik, yakni peserta didik, ukuran obyek yang ingin diketahui adalah hasil belajar, tetapi kita tidak memiliki alat ukur standar berupa alat ukur fisik (seperti meteran) yang bisa dijadikan patokan. Hal ini terjadi karena aspek-aspek psikologis lebih banyak bersifat laten atau tersembunyi, sehingga kita tidak dapat mengukur secara langsung obyek ukurnya. Kita hanya dapat melakukan pengukuran terhadap gejalanya saja. Jika kita mengukur panjang meja, maka ukuran panjang meja tersebut secara langsung dapat kita lihat dari penggunaan meteran. Tetapi hal yang sama tidak dapat kita lakukan ketika kita mengukur hasil belajar peserta didik, karena hasil belajar tersebut tidak tampak secara fisik dan secara langsung tidak teramati. Yang bisa kita lakukan adalah melakukan pengamatan terhadap gejalanya saja, dengan indikator-indikator tertentu. Dengan demikian, pengukuran aspek-aspek psikologis lebih bersifat diagnostik.

Pengukuran dapat dibedakan menjadi pengukuran langsung, dan pengukuran tidak langsung. Suatu pengukuran disebut pengukuran langsung jika ukuran pada obyek pengukurannya dapat dibandingkan secara langsung dengan alat ukurnya. Misalnya ketika kita mengukur panjang meja, maka kita dapat secara langsung membandingkan ukuran meja tersebut dengan alat ukur berupa meteran. Demikian

pula halnya ketika kita mengukur suhu, kita dapat secara langsung membandingkan ukuran suhu pada obyek dibandingkan dengan ukuran pada thermometer. Artinya, ukuran yang kita peroleh dari hasil pengukuran, benar-benar menggambarkan keadaan ukuran sebenarnya dari obyek, bukan ukuran dari gejala yang ditimbulkannya.

Hal berbeda akan kita temukan pada pengukuran tidak langsung. Suatu pengukuran disebut pengukuran tidak langsung, jika kita tidak dapat secara langsung mengukur ukuran dari obyek yang kita ukur. Umumnya yang dapat kita ukur hanya gejala dari ukuran tersebut, setelah obyek diberikan respon tertentu. Sebagai ilustrasi, ketika seorang dokter mendiagnosa penyakit yang diderita seseorang berdasarkan gejala yang timbul. Seseorang yang suhu tubuhnya di atas normal, kehilangan indera penciuman dan perasa, maka orang tersebut bisa diduga atau didiagnosa terkena penyakit Covid-19. Dalam hal ini, penyakit Covid-19 menyerang paru-paru, tetapi untuk menentukan jenis penyakit tersebut dokter cukup mendiagnosa gejala-gejalanya saja, tanpa harus melakukan pembedahan pada paru-paru pasien. Contoh lain adalah seseorang yang didiagnosa menderita penyakit tifus, penyakit infeksi pada usus halus. Dalam hal ini dokter kadangkala hanya mendignosa berdasarkan gejala-gejalanya saja seperti suhu tubuh yang turun naik pada masa tertentu, tidak secara langsung melakukan pengamatan terhadap usus halus pasien. Dalam dua contoh tersebut, hasil pengukuran didasarkan pada pengamatan terhadap gejala yang timbul, tanpa harus melakukan pengukuran terhadap obyek pengukuran. Pengukuran semacam ini disebut pengukuran tidak langsung.

Pengukuran aspek-aspek psikologis umumnya termasuk ke dalam pengukuran tidak langsung, diantaranya adalah pengukuran hasil belajar dalam dunia pendidikan. Artinya, kita sebagai pengukur atau evaluator tidak dapat mengukur ukuran hasil belajar peserta didik secara langsung, karena yang dapat kita ukur hanya gejala-gejalanya saja. Pada

pengukuran tidak langsung, pengukur hanya dapat mengukur gejala atau respon dari obyek yang diukur, kemudian memberi skor pada respon tersebut. Dalam hal ini, pengukur tidak memperoleh keputusan yang eksak tentang hasil pengukurannya, karena skor yang didapat hanya berdasarkan gejala yang timbul setelah diberi stimulus tertentu. Dalam pengukuran pendidikan, stimulus yang diberikan berbentuk instrumen atau alat ukur tertentu, baik berbentuk tes maupun non tes.

Hasil pengukuran umumnya dinyatakan secara kuantitatif menggunakan angka tertentu. Angka tersebut disebut skor hasil pengukuran. Dalam konteks penggunaan tes sebagai alat ukur hasil belajar, maka skor dapat diartikan sebagai angka yang melambangkan jumlah jawaban benar yang diperoleh dari suatu pengukuran (Aries, 2011). Skor juga dapat diSkor adalah gunaan tes angka dalam rentang tertentu, yang menyatakan hasil pengukuran. Rentang skor disesuaikan dengan kebutuhan pengukuran. Kita bisa menggunakan skor dalam rentang 0-100, rentang 60-100, rentang 0-10, dan sebagainya.

Skor hasil pengukuran tidak langsung masih mengandung ketidakpastian atau kesalahan. Naga (1992) [1], menyatakan bahwa skor hasil pengukuran pendidikan masih bersifat probalistik karena mengandung unsur kekeliruan. Dengan kata lain, skor hasil pengukuran hasil belajar terdiri dari skor sebenarnya (*True score*) dan skor kekeliruan (*Error*), yang dapat dilambangkan dalam persamaan berikut :

$X = T + E$

X = skor hasil pengukuran

T = skor sebenarnya

E = *error* atau kekeliruan

Dengan demikian, jika seorang peserta didik memperoleh skor 75 dari hasil sebuah tes hasil belajar, maka skor 75 tersebut belum tentu menggambarkan kemampuan

sebenarnya dari peserta didik. Banyak kombinasi skor T dan E yang mungkin terjadi, misalnya :

75 = 60 + 15……..(1)

75 = 95 + (-20) ……..(2).

Pada persamaan (1) di atas, kemampuan sebenarnya dari peserta didik adalah 60, tetapi karena terdapat skor kekeliruan sebesar 15 maka skor yang diperoleh peserta didik atau skor hasil pengamatannya adalah 75. Kemungkinan berbeda terjadi pada persamaan (2), yang mana kemampuan sebenarnya dari peserta didik adalah 95, tetapi karena terdapat skor kekeliruan sebesar -20 maka skor yang diperoleh peserta didik atau skor hasil pengamatannya adalah 75. Ini berarti, pada sebuah skor yang kita peroleh dari tes hasil belajar misalnya, terdapat tak terhingga banyaknya kemungkinan pasangan skor T dan skor E.

Dalam pengukuran hasil belajar, tantangan utama pengukur atau evaluator adalah meminimalkan skor kekeliruan atau *error*. Jika diusahakan nilai skor kekeliruan atau *error* mendekati nol ($E \approx 0$), maka persamaan $X = T + E$ akan mendekati $X = T + 0$, sehingga nilai skor hasil pengamatan akan hampir sama dengan nilai *True* skor. Dengan kata lain, jika dapat diusahakan $E \approx 0$, maka akan terjadi $X \approx T$. Artinya, dengan mengusahakan skor kekeliruan yang sekecil mungkin, maka skor hasil pengamatan yang kita peroleh akan mampu menggambarkan kemampuan sebenarnya dari peserta didik. Masalahnya adalah, skor X adalah hasil pengamatan, sehingga dapat kita amati skornya. Sedangkan skor T dan E tidak dapat kita amati, dan dapat dikendalikan secara teoretis. Untuk memperkecil *error* tersebut, dibutuhkan instrumen yang handal. Kehandalan instrumen tersebut akan kita bahas pada bagian berikutnya dalam buku ini.

B. Penilaian

Penilaian merupakan tahapan berikutnya setelah pengukuran. Penilaian adalah upaya untuk melakukan interpretasi atau menterjemahkan hasil pengukuran. Penilaian adalah interpretasi terhadap skor hasil pengukuran. Interpretasi tersebut dapat berupa pengolahan secara deskriptif, menggunakan analisis statistika, atau membandingkan skor tersebut dengan kriteria yang telah ditetapkan sebelumnya.

Hasil penilaian disebut sebagai nilai. Di sinilah letak perbedaan istilah skor dengan nilai, karena skor menyatakan hasil pengukuran, sedangkan nilai sudah merupakan pengolahan interpretasi terhadap skor. Skor dinyatakan dalam angka, tetapi nilai bisa dinyatakan dalam angka, huruf, kategori, dan sebagainya. Nilai merupakan ubahan dari skor yang diperoleh dari kerja analisis dan pendekatan tertentu (Aries, 2011). Misalnya dari pengukuran hasil belajar Matematika di kelas V SD menggunakan suatu tes, peserta didik A memperoleh skor hasil tes 75, dan peserta didik B memperoleh skor hasil tes 60. Skor tersebut kemudian akan diolah menjadi nilai dengan cara dibandingkan terhadap Kriteria Ketuntasan minimal (KKM) Matematika yang berlaku di sekolah tersebut, misalnya KKM=70. Skor 75 yang diperoleh peserta didik A dan B bisa dinyatakan menjadi nilai, yakni bahwa peserta didik A telah "tuntas" dan peserta didik B "belum tuntas". Angka 75 dan 60 adalah skor hasil pengukuran, KKM = 70 adalah kriteria penilaian, dan kata "tuntas:" dan "belum tuntas" adalah nilai.

Contoh lain tentang skor dan nilai adalah ketika seorang mahasiswa memperoleh skor tes ujian akhir semester. Mahasiswa X memperoleh skor 57, mahasiswa Y memperoleh skor 75, dan mahasiswa Z memperoleh skor 82. Kriteria penilaian yang digunakan sebagai acuan adalah sebagai berikut:

Tabel 2.1. Contoh kriteria penilaian

Interval skor	Nilai
00-39	E
40-55	D
56-69	C
70-79	B
80-100	A

Berdasarkan kriteria penilaian tersebut, maka mahasiswa X akan memperoleh nilai C, mahasiswa Y akan memperoleh nilai B, dan mahasiswa Z akan mendapatkan nilai A. dalam hal ini, angka-angka 57, 75 dan 82 adalah skor, sedangkan huruf A, B, dan C disebut nilai.

Penilaian juga dilakukan dengan mengacu pada dua pola acuan, yang disebut Penilaian Acuan Normatif (PAN), dan Penilaian Acuan Patokan (PAP). Kedua pola acuan tersebut akan dibahas pada bagian tersendiri dari buku ini.

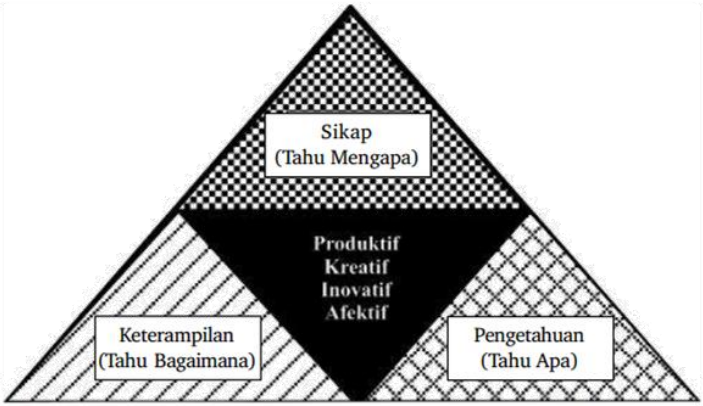
BAB III
ASPEK-ASPEK PENGUKURAN HASIL BELAJAR

Berdasarkan taksonomi yang digagas oleh Benjamin S. Bloom pada tahun 1956, hasil belajar dapat dibedakan menjadi 3 aspek atau ranah, yakni aspek kognitif (*cognitive),* sikap (*affective),* dan aspek keteranpilan (*psychomotor).* Taksonomi sendiri adalah ilmu yang mempelajari tentang pengelompokkan, struktur atau hirarki. Pengelompokkan ini kemudian dikenal luas dengan istilah taksonomi Bloom. Pengelompokkan tersebut dapat dijelaskan sebagai berikut:

1. Aspek kognitif, adalah aspek yang berisi pengertian, pengetahuan, pemahaman, dan keterampilan berfikir lainnya. Hasil belajar peserta didik yang berupa kemampuan mengingat, memahami, mengaplikasikan konsep, dan semacamnya, merupakan contoh hasil belajar aspek kognitif.
2. Aspek sikap, adalah hasil belajar terkait dengan perasaan, emosi, dan apresiasi. Contohnya adalah minat belajar, motivasi belajar, sikap terhadap mata pelajaran tertentu, kemampuan adaptasi, dan sebagainya.
3. Aspek keterampilan, adalah hasil belajar berbentuk perilaku motorik atau penguasaan keahlian tertentu. Misalnya adalah kemampuan mengetik, kemampuan dan kecepatan berlari dalam mata pelajaran olahraga, mengoperasikan mesin, dan sebagainya.

Dalam konteks proses pembelajaran, ketiga aspek hasil belajar tersebut tidaklah berdiri sendiri. Ada saling

keterkaitan antara satu aspek dengan aspek lainnya. Asrul dkk (2015), memandang keterkaitan antara ketiga aspek tersebut sebagai penunjang semua *sofskill* dan *hardskill* hasil belajar yang diharapkan, yang membentuk kemampuan (*ability)* seseorang. Hal itu digambarkan dalam piramida hasil belajar sebagai berikut:



Gambar 3.1. Keterkaitan antara aspek-aspek hasil belajar

Dengan demikian, aspek kognitif menekankan pada peserta didik untuk mengetahui tentang sesuatu, aspek sikap menekankan pada peserta didik untuk mengetahui tentang mengapa, dan aspek keterampilan menekankan pada peserta didik untuk mengetahui tentang bagaimana (Asrul dkk, 2015). Pada kenyataannya, aspek pengetahuan juga mendasari adanya sikap dan keterampilan. Demikian pula sebaliknya, tumbuhnya aspek sikap positif terhadap sesuatu akan menumbuhkan motivasi untuk meningkatkan pengetahuan dan keterampilan. Dengan kata lain, aspek-aspek hasil belajar tersebut bukanlah berdiri masing-masing, tetapi ada keterkaitan berupa hubungan timbal balik antara ketiga aspek hasil belajar tersebut. Dalam pandangan para ahli pengukuran, kemampuan (*ability*) diartikan sebagai kombinasi antara pengetahuan, sikap, dan keterampilan. Hal

ini misalnya tergambar dalam beberapa pandangan Bloom, Englehard, Furst, dan Krathwohl (Lestari & Setiawan, 2017). Pandangan ini akan berpengaruh terhadap cara kita mengukur ketiga aspek tersebut.

A. Aspek Hasil Belajar

Bloom juga merinci aspek-aspek hasil belajar, yakni pengetahuan, sikap dan keterampilan tersebut menjadi aspek-aspek yang lebih operasional, yakni:

1. Aspek pengetahuan (*cognitive)*, dibagi menjadi 6 sub aspek atau tahapan kemampuan berfikir, yakni:
    a. Ingatan (C1)
    b. Pemahaman (C2)
    c. Aplikasi (C3)
    d. Analisis (C4)
    e. Sintesis (C5).
    f. Evaluasi (C6).
2. Aspek sikap (*affective)*, dibagi menjadi 5 sub aspek atau tahapan bersikap, yakni:
    a. Menerima (A1)
    b. Menanggapi (A2)
    c. Menghargai (A3)
    d. Mengorganisir (A4)
    e. Karakterisasi (A5)
3. Aspek keterampilan (*psychomotor)*, dibagi menjadi 4 sub aspek atau tahapan berperilaku berikut:
    a. Menirukan (P1)
    b. Memanipulasi (P2)
    c. Mengalamiahkan (P3)
    d. Mengartikulasikan (P4)

Taksonomi Bloom merupakan dasar pengelompokkan hasil belajar yang masih dipakai hingga saat ini, meskipun telah mengalami beberapa revisi. Revisi antara lain dikemukakan oleh Anderson & Krathwohl (2001). Revisi ini dipublikasikan pada tahun 2001. Taksonomi Bloom revisi ini hanya dilakukan terhadap aspek kognitif saja, dengan merevisi penggunaan kata operasional, yakni berubah dari kata benda menjadi kata kerja (misalkan kemampuan aplikasi berubah menjadi mengaplikasikan). Sebagai contoh, dalam versi taksonomi Bloom yang direvisi, ranah C1 yang merupakan aspek pengetahuan, direvisi menjadi aspek kemampuan mengingat. Sedangkan ranah C5 (sintesis) dalam taksonomi Bloom yang lama, direvisi dan bertukar posisinya dengan aspek C6 (evaluasi), sehingga susunannya menjadi C5 (evaluasi) dan C6 (sintesis) dalam taksonomi Bloom yang baru. Urutan ranah ini menggambarkan urutan kemampuan berfikir kognitif. Gambaran tentang revisi taksonomi Bloom pada aspek kognitif adalah sebagai berikut:

Tabel 3.1: Perubahan aspek pada taksonomi Bloom

Aspek taksonomi Bloom lama	Menjadi	Aspek taksonomi Bloom revisi
C1 : Pengetahuan	→	C1 : Mengingat
C2 : Pemahaman	→	C2 : Memahami
C3 : Aplikasi	→	C3 : Mengaplikasikan
C4 : Analisis	→	C4 : Menganalisis
C5 : Sintesis	→	C5 : Mengevaluasi
C6 : Evaluasi	→	C6 : Mencipta/mensintesis

Selain adanya perubahan kata, revisi oleh Anderson dan Krathwohl tersebut juga dilakukan dengan menambahkan dimensi pengukuran pada aspek kognitif, yakni dimensi

pengetahuan kognitif dan dimensi proses kognitif (Anderson, 2010). Penambahan dimensi kognitif ini dianggap penting untuk meningkatkan akurasi hasil pengukuran, karena bukan hanya mengukur hasil tetapi juga mengukur proses berfikir.

Jika pada taksonomi Bloom yang lama, hasil belajar aspek kognitif hanya diwakili oleh aspek dimensi proses kognitif semata, yakni dari aspek C1 hingga C6, maka dalam taksonomi Bloom versi revisi, dimensi pengetahuan kognitif juga dianggap sebagai hasil belajar. Dimensi pengetahuan kognitif tersebut dibedakan menjadi 4 kategori, yakni pengetahuan faktual, pengetahuan konseptual, pengetahuan prosedural, dan pengetahuan metakognitif. Dengan demikian, pada taksonomi Bloom versi revisi, aspek-aspek kognitif C1 hingga C6 dapat dikelompokkan menggunakan matriks ukuran 4x6 berdasarkan 4 jenis dimensi pengetahuan kognitif dan 6 aspek proses kognitif. Hal ini lebih memperkaya tujuan pembelajaran dan memperkaya tujuan pengukuran. Model matriks inilah yang sekarang banyak digunakan sebagai dasar dalam perumusan tujuan pembelajaran dengan pendekatan taksonomi Bloom versi baru. Gambaran tentang tujuan pembelajaran berdasarkan taksonomi Bloom versi revisi adalah sebagai berikut:

Tabel 3.2: Matriks tujuan pembelajaran berdasarkan taksonomi Bloom revisi

	C1 Mengingat	C2 Memahami	C3 Mengaplikasi	C4 Menganalisis	C5 Mengevaluasi	C6 Mencipta
Pengetahuan faktual	C1 Faktual	C2 faktual	C3 faktual	C4 faktual	C5 faktual	C6 faktual
Pengetahuan konseptual	C1 konseptual	C2 konseptual	C3 konseptual	C4 konseptual	C5 konseptual	C6 konseptual
Pengetahuan Procedural	C1 prosedural	C2 prosedural	C3 prosedural	C4 prosedural	C5 prosedural	C6 prosedural
Pengetahuan metakognitif	C1 metakognitif	C2 metakognitif	C3 metakognitif	C4 metakognitif	C5 metakognitif	C6 metakognitif

Pengayaan dalam taksonomi Bloom revisi, memungkinkan guru untuk merinci aspek pengetahuan jenis apa yang diharapkan sebagai hasil belajar, sehingga hal itu juga memungkinkan pengukur atau evaluator dapat merinci lebih

spesifik aspek pengetahuan yang ingin diukur melalui evaluasi pembelajaran. Dengan memetakan dimensi pengetahuan kognitif, maka aspek hasil belajar kognitif dapat dipetakan secara lebih rinci, dengan berpedoman pada batasan berikut:

1. Dimensi pengetahuan faktual (*factual knowledge dimension*)
    a. Pengetahuan tentang terminologi, istilah.
    b. Pengetahuan tentang detail dan unsur-unsur kejadian, subyek, waktu, keadaan, dan semacamnya.
2. Dimensi pengetahuan konseptual (*conceptual knowledge dimension*)
    a. Pengetahuan tentang klasifikasi,
    b. Pengetahuan tentang prinsip
    c. Pengetahuan tentang generalisasi
    d. Pengetahuan tentang model, struktur, dan teori.
3. Dimensi pengetahuan prosedural (*procedural knowledge dimension*)
    a. Pengetahuan tentang algoritma
    b. Pengetahuan tentang teknik, metode, dan keterampilan pada bidang tertentu
    c. Pengetahuan tentang penggunaan suatu prosedur secara tepat.
4. Dimensi pengetahuan metakognitif (*meta-cognitive knowledge dimension*)
    a. Pengetahuan tentang strategi
    b. Pengetahuan tentang konteks dan kondisi
    c. Pengetahuan tentang diri sendiri

Revisi dalam taksonomi Bloom juga membawa dampak terhadap aspek aspek hasil belajar yang diukur. Misalnya

terdapat perbedaan antara pengetahuan atau mengetahui sesuatu, dengan mengetahui tentang cara melakukan sesuatu (Mulatsih, 2021). Adanya pemetaan secara rinci pengetahuan kognitif, menyebabkan perbedaan makna antara keduanya, yang juga beriakibat terhadap cara pengukurannya. Mengetahui sesuatu dapat berupa pengetahuan faktual, konseptual, atau metakognitif. Sedangkan mengetahui tentang cara melakukan sesuatu termasuk pengetahuan procedural. Pemetaan secara lebih rinci aspek kognitif, membantu evaluator menyusun kata-kata operasional yang lebih terukur.

Adanya revisi pada aspek kognitif, menyebabkan adanya perubahan pada kata-kata operasional yang biasa digunakan dalam merumuskan tujuan pembelajaran maupun tujuan pengukuran hasil belajar. Pemilihan kata yang tepat dan operasional dalam tujuan pembelajaran dan tujuan pengukuran, akan memudahkan guru untuk mengindentfikasi aspek hasil belajar yang akan diukur. Selain itu, juga akan memudahkan guru selaku evaluator dalam menentukan kata atau kalimat perintah dalam menyusun butir-butir pertanyaan atau pernyataan pada instrument yang digunakan.

Deskripsi tentang kata-kata operasional yang digunakan dalam menentukan tujuan pembelajaran adalah sebagai berikut:

B. Kata Operasional untuk Aspek Kognitif

Dalam menyusun tujuan pembelajaran dan tujuan pengukuran aspek kognitif, guru harus terlebih dahulu menentukan aspek proses dan dimensi kognitif seperti apa yang dihatapkan akan terjadi ada peserta didik setelah memperoleh pembelajaran. Misalkan hasil pembelajaran yang diharapkan adalah: Peserta didik dapat **menyebutkan** nama presiden RI pertama, maka kata **menyebutkan** dalam tujuan pembelajaran tersebut merupakan kemampuan yang diharapkan terjadi sebagai hasil pembelajaran. Dengan

demikian, guru akan memilih untuk menggunakan kata **menyebutkan** sebagai kata operasional dalam tujuan pembelajaran maupun tujuan pengukuran.

Gambaran tentang kata operasional yang dapat digunakan guru dan evaluator pada hasil belajar kognitif adalah sebagai berikut:

Tabel 3.3.: Kata operasional pada aspek kognitif

Mengingat (C1)	Memahami (C2)	Mengaplikasi (C3)	Menganalisis (C4)	Mengevaluasi (C5)	Mencipta (C6)
Mengutip	Menjelaskan	Mengurutkan	Menganalisis	Membandingkan	Mengabstraksi
Menyebutkan	Mengkategori	Menentukan	Mengaudit	Menyimpulkan	Mengatur
Menjelaskan	Merinci	Menerapkan	Memecahkan	Menilai	Menganimasi
Menggambar	Mengasosiasi	Menyesuaikan	Menegaskan	Mengarahkan	Mengkode
Membilang	Membedakan	Mengkalkulasi	Mendeteksi	Mengkritik	Menyusun
Mengidentifikasi	Menghitung	Memodifikasi	Mendiagnosis	Menimbang	Mengarang
Mendaftar	Mengkontraskan	Menghitung	Menyeleksi	Memutuskan	Membangun
Menunjukkan	Mengubah	Menggambarkan	Merinci	Memisahkan	Menanggulangi
Memberi label	Menguraikan	Menggunakan	Menominasikan	Memperjelas	Menghubungkan
Memberi indek	Menjalin	Menilai	Mendiagramkan	Menegaskan	Menciptakan
Memasangkan	Menggali	Melatih	Mengkorelasikan	Menafsirkan	Mengkreasikan
Memberi nama	Mencontohkan	Mengemukakan	Merasionalkan	Mempertahankan	Merancang
Menandai	Menerangkan	Mengadaptasi	Menguj	Merangkum	Merencanakan
Membaca	Menyimpulkan	Menyelidiki	Menjelajah	Membuktikan	Meningkatkan
Menghafal	Meramalkan	Mengoperasikan	Membagankan	Memvalidasi	Memfasilitasi
Meniru	Merangkum	Mempersoalkan	Menyimpulkan	Mengetes	Membentuk
Mencatat	Menjabarkan	Mengkonsepkan	Menelaah	Mendukung	Merumuskan
Mengulang		Melaksanakan	Memaksimalkan	Memilih	Menggeneralisasi
Memilih		Memproduksi	Memerintahkan	Memproyeksika	Menggabungkan
Menyatakan		Memproses	Mengedit		Mereparasi
Mentabulasi		Mengaitkan	Mengaitkan		Menyiapkan
Memberi kode		Menyusun	Memilih		Memproduksi
Menuliskan kembali		Mensimulasikan	Mengukur		Merangkum
		Memecahkan	Melatih		Merekonstruksi
		Mentabulasi	Mentransfer		Membuat

C. Kata Operasional untuk Aspek Afektif

Kata-kata operasional yang dapat digunakan dalam aspek afektif adalah:

Tabel 3.4. Kata operasional pada aspek afektif

Menerima (A1)	Menanggapi (A2)	Menghargai (A3)	Mengorganisir (A4)	Karakterisasi (A5)
Memilih	Menjawab	Mengasumsikan	Menganut	Mengubah
Mempertanyakan	Membantu	Meyakini	Mengubah	perilaku
Mengikuti	Mengajukan	Melengkapi	Menata	Mempengaruhi
Memberi	Mengompromikan	Meyakinkan	Mengklasifikasi	Mendengarkan
Menganut	Menyenangi	Memperjelas	Mengombinasi	Mengkualifikasi
Mematuhi	Menyambut	Memprakarsai	Mempertahankan	Melayani
Meminati	Mendukung	Mengimani	Membangun	Menunjukkan
		Mengundang	Membentuk	Membuktikan
		Menggabungkan	Memadukan	Memecahkan
		Mengusulkan	Mengelola	
		Menekankan	Menegosisasi	
		Menyumbang	Merembuk	
		Mempraktikkan		

D. Kata Operasional untuk Aspek Psikomotor

Kata operasional yang dapat digunakan dalam tujuan pembelajaran dan tujuan pengukuran aspek keterampilan adalah sebagai berikut:

Tabel 3.5 : Kata operasional pada aspek psikomotor

Menirukan (P1)	Memanipulasi (P2)	Mengalamiahkan (P3)	Mengartikulasi (P4)
Mengaktifkan	Mengoreksi	Mengalihkan	Mengalihkan
Menyesuaikan	Mendemontrasikan	Menggantikan	Mempertajam
Menggabungkan	Merancang	Memutar	Membentuk
Melamar	Memilah	Mengirim	Memadankan
Mengatur	Melatih	Memindahkan	Menggunakan
Mengumpulkan	Memperbaiki	Mendorong	Memulai
Menimbang	Mengidentifikasi	Menarik	Menyetir
Memperkecil	Mengisi	Memproduksi	Menjeniskan
Membangun	Menempatkan	Mengoperasikan	Menempel
Mengubah	Membuat	Mengemas	Mensketsa
Membersihkan	Memanipulasi	Membungkus	Melonggarkan
Memposisikan	Mencampur		Menimbang
Mengonstruksi			

BAB IV
INSTRUMEN PENGUKURAN

A. Makna Instrumen Pengukuran

Sebagaimana telah dijelaskan pada bagian sebelumnya, pengukuran adalah membandingkan ukuran suatu obyek yang diukur, dengan menggunakan alat ukur standar yang telah disepakati/diketahui kehandalannya dan digunakan sebagai patokan. Pengukuran merupakan penentuan suatu angka bagi suatu obyek secara sistematis guna menggambarkan karakteristik obyek tersebut (Mardapi, 2012). Dengan demikian, pengukuran tidak dapat dipisahkan dari alat ukur atau instrumen pengukuran.

Instrumen pengukuran adalah alat standar yang digunakan sebagai alat ukur, sehingga dari alat tersebut diperoleh angka atau skor yang bisa dipercaya. Skor yang dihasilkan oleh suatu instrumen pengukuran hasil belajar, merupakan angka yang dianggap bisa mewakili karakteristik dan indikator hasil belajar yang telah diukur, selama instrumen yang digunakan adalah instrumen yang standar dan dapat dipercaya. Misalnya angka 80 yang diperoleh seorang peserta didik dari hasil tes Matematika, melambangkan penguasaan dan hasil belajarnya pada mata pelajaran Matematika tersebut.

Suatu instrumen disebut sebagai alat ukur standar jika instrumen itu telah diyakini akurasinya dan disepakati secara luas sebagai alat ukur. Misalkan meteran adalah alat ukur standar untuk mengukur panjang benda, karena diyakini mampu mengukur panjang benda secara akurat dan telah

disepakati secara luas. Timbangan juga merupakan alat ukur yang standar untuk mengukur berat benda, karena telah diyakini akurasinya dan disepakati secara luas.

Dalam pengukuran hasil belajar, aspek yang kita ukur merupakan aspek psikologis yang bersifat laten atau tersembunyi, sehingga yang dapat kita ukur hanyalah gejalanya saja. Peserta pengukuran diberikan suatu stimulus atau rangsangan tertentu, melalui suatu instrumen, sehingga dihasilkan respon. Respon itulah yang kemudian dicatat dalam bentuk angka. Mardapi (2012) menyatakan bahwa pengukuran tidak lagsung pada dasarnya merupakan upaya kuantifakasi terhadap obyek non fisik, termasuk diantaranya adalah gejala. Pengukuran aspek-aspek psikologis, termasuk hasil belajar, merupakan pengukuran tidak langsung. Tantangan utama pengukuran tidak langsung adalah menyiapkan alat ukur yang standar, sehingga hasil pengukurannya dapat dipercaya. Artinya, jika instrumen yang digunakan kurang handal, maka skor hasil pengukurannya kurang dapat dipercaya. Masalahnya adalah, dalam pengukuran hasil belajar, kita sering mengalami kesulitan untuk menemukan suatu instrumen yang standar dan dapat dipercaya pada suatu waktu tertentu. Soal-soal pada bank soal sering telah *out of date.* Penggunaan butir-butir soal tes dari bank soal misalnya, sering terkendala dengan adanya perubahan karakteristik peserta didik dan perubahan kurikulum pendidikan, sehingga soal-soal tersebut kurang relevan.

Dengan demikian, tantangan utama dalam pengukuran hasil belajar adalah menyusun instrumen yang *up to date*, yang mampu mengukur karateristik dan capaian hasil belajar peserta didik pada suatu waktu tertentu. Untuk itu diperlukan perencanaan yang matang, kehati-hatian dan ketelitian dalam menyusun instrumen pengukuran hasil belajar, sehingga diperoleh instrumen yang standar dan hasil pengukurannya dapat dipercaya. Langkah-langkah penyusunan instrumen

yang baik harus diikuti agar diperoleh instrumen pengukuran hasil belajar yang handal.

Sebagai catatan, perlu dibedakan makna antara instrumen dengan butir-butir instrumen. Instrumen bermakna sebagai kumpulan atau seperangkat alat, yakni merupakan kumpulan butir-butir atau item pertanyaan atau pernyataan. Misalnya kita akan mengukur hasil belajar IPS menggunakan tes hasil belajar IPS yang terdiri dari 25 butir soal pilihan ganda (PG) dan 5 soal utaian. Dalam konteks ini, tes hasil belajar IPS merupakan perangkat atau instrumen pengukuran, sedangkan butir-butir pertanyaan dalam tes tersebut merupakan butirbutir instrumen. Pemahaman tentang perbedaan makna tersebut penting untuk difahami terutama dalam pembahasan tentang langkah-langkah penyusunan instrumen dan pengujian kehandalan instrumen.

B. Jenis Instrumen Pengukuran Hasil Belajar.

Secara umum, instrumen pengukuran hasil belajar dapat dibedakan menjadi jenis tes dan non tes (Lestari & Setiawan, 2017). Pembagian ini didasarkan pada teknik penggunaannya. Perbedaan antara keduanya dijelaskan sebagai berikut:

1. Tes

Pengertian tes dapat dipandang dari dua sisi, yakni tes sebagai teknik pengukuran dan tes sebagai alat ukur. Sebagai teknik pengukuran, tes (atau disebut pula teknik tes) didefinisikan sebagai suatu prosedur yang digunakan untuk mengukur hasil belajar, khususnya dalam aspek kognitif dan psikomotor. Sedangkan sebagai alat ukur pengukuran pendidikan, tes didefinisikan sebagai sekumpulan pertanyaan atau pernyataan yang digunakan untuk mengukur hasil belajar. Ciri utama tes sebagai alat ukur dapat dilihat dari pola jawaban peserta didik sebagai orang yang memberi respon (responden), yang mana jawaban responden bisa bernilai benar bisa pula bernilai salah. Artinya, ketika peserta didik diberikan suatu pertanyaan, maka jawaban yang diberikannya bisa

benar atau bisa pula salah. Hal ini merupakan ciri khas tes sebagai alat ukur. Misalkan ketika seorang peserta didik diberikan pertanyaan yang berbunyi "Siapakah nama Presiden RI yang pertama?", maka pertanyaan tersebut merupakan tes, karena jawaban peserta didik bisa bernilai benar (jika menjawab Ir. Sukarno), dan bisa pula bernilai salah (jika peserta didik menjawab selain Ir. Sukarno). Contoh lainnya adalah, ketika peserta didik diberikan pertanyaan "Berapakah jumlah dari 3+4 ?", maka pertanyaan tersebut merupakan tes karena jawaban peserta didik bisa bernilai benar (jika menjawab 7) atau bisa pula bernilai salah (jika menjawab selain 7).

Umumnya hasil tes dinyatakan secara kuantitatif dalam bentuk skor, walaupun kemudian skor tersebut pada akhirnya bisa dinyatakan dalam kategori-kategori kualitatif. Pemberian skor pada tes relatif lebih gampang, yakni dengan cara memberikan skor lebih tinggi pada jawaban benar dan skor lebih rendah pada jawaban salah. Misalkan pada tes pilihan ganda dan isian yang memiliki jawaban lebih pasti, maka kita dapat memberikan skor 1 pada setiap jawaban benar dan skor 0 pada jawaban salah. Pada tes berbentuk uraian yang nilai kebenaran jawaban peserta didik memiliki unsur relativitas, maka pemberian skor dapat menggunakan rentang tertentu, misalkan jawaban benar sempurna diberi skor 10, benar sebagian besar diberi skor 7, benar sebagian diberi skor 5, sama sekali tidak ada yang benar diberi skor 0.

2. Non Tes

Pengertian non tes juga dapat dipandang dari dua sisi, yakni non tes sebagai teknik pengukuran dan non tes sebagai alat ukur. Sebagai teknik pengukuran, non tes (atau disebut pula teknik non tes) didefinisikan sebagai suatu prosedur yang digunakan untuk mengukur hasil belajar non kognitif, khususnya adalah aspek afektif. Misalkan dalam pengukuran pendapat, sikap, motivasi,

minat, aspirasi, dan sebagainya. Sedangkan sebagai alat ukur pengukuran pendidikan, non tes didefinisikan sebagai sekumpulan pertanyaan atau pernyataan yang digunakan untuk mengukur hasil belajar afektif. Berbeda dengan tes, maka ciri utama non tes sebagai alat ukur adalah jawaban responden selalu bernilai benar atau dianggap bernilai benar. Sebagai contoh, ketika seorang peserta didik diberikan pertanyaan "Apakah saudara menyenangi mata pelajaran Matematika?", maka jawaban peserta didik akan selalu bernilai benar sekalipun mereka ada yang menjawab "sangat menyukai", "tidak menyukai", "sangat membenci", dan sebagainya. Contoh lain adalah pola respon peserta didik terhadap pertanyaan "Sebutkan cita-cita kamu", maka jawaban peserta didik yang mungkin sangat beragam harus dianggap sebagai jawaban yang bernilai benar.

Hasil pengukuran menggunakan non tes, dapat dinyatakan secara kualitatif maupun kuantitatif dalam bentuk skor tertentu yang menyatakan suatu kecenderungan. Cara pemberian skor terhadap instrumen non tes umumnya menggunakan skala tertentu yang dibuat bergradasi. Misalkan untuk contoh pertanyaan "Apakah saudara menyenangi mata pelajaran Matematika?", yang digunakan untuk mengukur tingkat atau kecenderungan minat peserta didik terhadap mata pelajaran Matematika. Jika kita ingin mengukur kecenderungan minat yang positif, maka jawaban "sangat menyukai" bisa diberikan skor 5, "menyukai" diberi skor 4, "Ragu-ragu" diberi skor 3, "tidak menyukai" diberi skor 2, dan "sangat tidak menyukai" diberi skor 1. Sebaliknya, jika kita ingin mengukur kecenderungan minat yang negative (misalkan tingkat ketidaksukaan terhadap mata pelajaran Matematika), maka kita bisa menggunakan pola pemberian skor sebaliknya, yakni "sangat menyukai" bisa diberikan skor 1, "menyukai" diberi skor 2, "Ragu-ragu" diberi skor 3, "tidak menyukai" diberi skor 4, dan "sangat

tidak menyukai" diberi skor 5. Terdapat beberapa pendapat yang menyarankan agar pilihan jawaban yang bersifat ambigu atau ragu-ragu sebaiknya dihilangkan. Jika pendapat tersebut yang kita gunakan, maka kita dapat memperkecil rentang pemberian skor dari angka 1 sampai 4 sesuai dengan jumlah alternatif jawaban.

C. Bentuk-Bentuk Instrumen Pengukuran Hasil Belajar

1. Bentuk-Bentuk Tes

Sebagai instrumen pengukuran, tes dapat dibedakan menjadi tes obyektif dan tes subyektif. Penggunaan istilah tes subyektif bukan berarti bahwa tes tersebut tidak obyektif, tetapi lebih bersifat pada pola jawaban bebas yang diberikan oleh responden. Artinya, setiap tes semestinya harus obyektif, baik dalam penggunaan maupun pemberian skornya.

Tes juga dapat dibedakan menjadi (1). Tes tertulis, (2). Tes lisan, (3). Tes pebuatan atau kinerja. Pembagian ini didasarkan pada teknik penggunaannya. Tes tertulis merupakan seperangkat pertanyaan yang diberikan secara tertulis kepada peserta tes, yang kemudian harus dijawab secara tertulis pula. Tes lisan adalah sejumlah pertanyaan yang diberikan secara lisan atau tertulis, tetapi harus dijawab secara lisan oleh peserta tes. Sedangkan tes praktik adalah sejumlah pertanyaan atau perintah yang diberikan secara tertulis atau lisan, yang dijawab secara praktik oleh peserta tes.

Tes tertulis dapat dibedakan berdasarkan bentuk pertanyaan dan pola jawaban respsondennya, yakni sebagai berikut:

a. Tes pilihan ganda (PG)

Tes pilihan ganda adalah tes yang terdiri dari pertanyaan/pernyataan dan beberapa pilihan jawaban yang disiapkan, sehingga peserta tes (responden) tinggal memilih salah satu alternatif

jawaban tersebut. Tes jenis ini terdiri dari batang tubuh soal (disebut juga stem soal), dan kelompok alternatif jawaban. Alternatif jawaban pada soal PG terdiri dari satu jawaban benar, dan sisanya berfungsi sebagai pengecoh (distraktor). Tes jenis ini disusun dengan terlebih dahulu membuat suatu kalimat pentanyaan/pernyataan secara utuh yang memiliki suatu kata kunci, kemudian kata kunci tersebut dihilangkan dalam stem soal dan dijadikan sebagai salah satu alternatif jawaban. Misalkan kalimat utuhnya adalah "Nama Presoden RI pertama adalah Ir. Sukarno". Salah satu kata kuncinya adalah Ir. Sukarno, maka kata kunci tersebut kemudian dihilangkan dan dijadikan alternatif jawaban. Akhirnya stem soal dan alternatif jawabannya adalah:

Nama Presiden RI pertama adalah:

a. Ir. Sukarno

b. Suharto

c. B.J. Habibie

d. Abdurahman Wahid

e. Susilo Bambang Yudhoyono

Contoh soal di atas adalah tes berbentuk pilihan ganda (PG) dengan alternatif jawaban sebanyak 5 pilihan. Jumlah alternatif pilihan dapat disesuaikan dengan tujuan pengukuran, kebutuhan, dan karakteristik peserta tes.

Kelebihan penggunaan tes PG antara lain adalah:

- Mampu mencakup materi pengukuran yang lebih luas
- Proses koreksi lebih mudah dan relatif cepat
- Lebih mudah dikerjakan oleh peserta tes.

- Memiliki nilai kebenaran lebih pasti karena hanya ada satu jawaban benar.
- Pengembangan variasi soal lebih mudah. Evaluator dapat menyusun beberapa variasi butir soal untuk indikator yang sama.
- Lebih mudah digunakan untuk evaluasi model daring.

Sedangkan kelemahan penggunaan tes PG yang menonjol adalah:

- Hanya cocok untuk mengukur aspek-aspek hasil belajar kognitif yang sederhana seperti ingatan (C1) dan pemahaman (C2).
- Relatif membutuhkan waktu yang lebih lama untuk menyusunnya.
- Adanya kesulitan menyusun pengecoh (distraktor) yang homogen atau mirip dengan alternatif jawaban.
- Memungkinkan peserta tes menjawab soal secara tebakan atau *guessing.*
- Peserta tes tidak diberikan kebebasan untuk mengemukakan ide, kritik, dan pendapat mendalam tentang materi yang diukur, karena alternatif jawaban dibatasi.
- Evaluator tidak dapat mengetahui proses berfikir peserta tes,

Beberapa pedoman dalam menyusun soal PG adalah:

- Susunlah stem soal sesuai dengan tujuan pengukuran dan kemampuan peserta tes,
- Usahakan kalimat dalam stem soal adalah kalimat singkat, tegas, jelas, dan tidak bertele-tele, serta hanya menanyakan tentang satu pokok pertanyaan saja.

- Buatlah alternatif dan kunci jawaban serta pedoman pemberian skornya segera setelah stem soal dibuat.
- Susunlah alternatif jawaban soal PG secara homogen dan mirip satu sama lain. Hal ini penting agar pengecoh (disktraktor) berfungsi.

b. Tes Isian Singkat

Tes isian singkat adalah tes yang terdiri dari pertanyaan atau pernyataan (stem soal) dan bagian jawaban soal yang dikosongkan, biasanya diberi tanda titik-titik kosong. Bentuk tes isian yang umum adalah berupa melengkapi kalimat, walaupun ada juga berbentuk kalimat perintah dan pertanyaan. Bagian yang dihilangkan dan merupakan jawaban tersebut dapat berupa kata, frase, nama tempat, nama tokoh, atau semacamnya yang bersifat pasti.

Aspek kognitif yang diukur biasanya mencakup kemampuan mengingat dan memahami. Kemampuan tersebut antara lain mencakup kemampuan menyebutkan istilah, menyebutkan fakta, menyebutkan prinsip, menyebutkan prosedur, menginterpretasi data secara sederhana, melengkapi kalimat atau persamaan, dan semacamnya.

Serupa dengan tes PG, soal isian untuk tes isian singkatnya adalah sebuah pernyataan utuh yang bagian kata pentingnya kemudian dihilangkan dan diganti dengan bagian kosong atau titik-titik yang harus diisi peserta tes. Misalkan kalimat utuhnya berbunyi: “Ibukota provinsi Kalimantan Tengah adalah Palangka Raya”, maka beberapa kata pentingnya adalah kata “Kalimantan Tengah” dan “Palangka Raya”. Dengan demikian, jika kalimat pernyataan itu akan diubah menjadi butir soal isian, maka salah satu kata pentingnya bisa dihilangkan/dikosongkan dan diganti menjadi titik-

titik. Bunyi butir soalnya menjadi : “Ibukota provinsi Kalimantan Tengah adalah ……..”. Perlu diingat bahwa dalam menyusun soal isian, penghapusan kata penting dalam kalimat jangan sampai mengaburkan makna pernyataan sehingga peserta tes kesulitan mengisi bagian yang dikosongkan.

Kelebihan tes isian singkat antara lain :

- Stem soal dan jawaban yang singkat akan memudahkan evaluator untuk menyusun butir soal.
- Materi pengukuran dapat mencakup materi pembelajaran yang luas.
- Lebih mudah dikerjakan oleh peserta tes.
- Mengurangi kemungkinan jawaban tebakan atau *guessing.*
- Proses koreksi dan pemberian skor lebih mudah.
- Memiliki nilai kebenaran lebih pasti karena hanya ada satu jawaban benar.
- Dapat digunakan untuk evaluasi model daring.

Sedangkan kelemahan tes isian singkat adalah :

- Sulit digunakan untuk mengukur aspek-aspek hasil belajar kognitif yang kompleks seperti C4, C5 dan C6.
- Peserta tes tidak diberikan kebebasan untuk mengemukakan ide dan pendapat mendalam tentang materi yang diukur, karena terbatasnya kalimat jawaban.
- Evaluator tidak dapat mengetahui proses berfikir peserta tes,

Beberapa kaidah yang harus dipenuhi saat penyusunan butir soal isian singkat antara lain:

- Susunlah stem soal berdasarkan tujuan pengukuran yang telah ditetapkan dan aspek hasil belajar kognitif yang akan diukur.
- Rumusan stem soal menggunakan kalimat yang singkat, tegas, jelas, dan hanya mengandung satu pokok masalah, sehingga mudah difahami peserta tes.
- Hindari rumusan stem soal yang mengarah pada kunci jawaban.
- Susunlah kunci jawaban segera setelah menyusun stem soal
- Kunci jawaban merupakan kata atau kalimat yang pasti.

c. Tes Menjodohkan

Tes menjodohkan adalah tes yang terdiri dari dua bagian atau dua kolom, yang mana peserta tes diminta menghubungkan, menjodohkan, mencocokkan, atau menyesuaikan antara pernyataan pada kolom pertama dengan jawaban pada kolom kedua. Biasanya pernyataan/pertanyaan yang merupakan stem soal diletakkan di kolom pertama (kolom kiri) dan pilihan jawaban diletakkan di kolom kedua (kolom kanan).

Tes menjodohkan umumnya digunakan untuk mengukur aspek-aspek kognitif yang bersifat hubungan sederhana, asosiatif, atau relasi antar fakta.

Kelebihan tes menjodohkan antara lain:

- Materi pengukuran dapat mencakup materi pembelajaran yang luas.
- Lebih mudah dikerjakan oleh peserta tes.
- Proses koreksi dan pemberian skor cukup mudah.

- Memiliki nilai kebenaran lebih pasti karena hanya ada satu jawaban benar.

Sedangkan kelemahan tes menjodohkan adalah :

- Sulit digunakan untuk mengukur aspek-aspek hasil belajar kognitif yang kompleks seperti C4, C5 dan C6.
- Peserta tes tidak diberikan kebebasan untuk mengemukakan ide dan pendapat mendalam tentang materi yang diukur, karena terbatasnya kalimat jawaban.
- Evaluator tidak dapat mengetahui proses berfikir peserta tes,

Beberapa kaidah yang harus dipenuhi saat penyusunan butir soal tes menjodohkan antara lain:

- Susunlah stem soal berdasarkan tujuan pengukuran yang telah ditetapkan dan aspek hasil belajar kognitif yang akan diukur.
- Akan sangat baik jika rumusan stem soal mengukur aspek yang homogen, sehingga alternatif jawaban juga homogen.
- Rumusan stem soal diletakkan di kolom kiri menggunakan kalimat yang singkat, tegas, jelas, dan hanya mengandung satu pokok masalah, sehingga mudah difahami peserta tes.
- Susunlah kunci jawaban segera setelah menyusun stem soal yang diletakkan di kolom kanan beserta pedoman penskorannya.
- Usahakan kunci jawaban merupakan satu kata yang pasti.
- Usahakan jumlah stem soal lebih sedikit dibandingkan jumlah alternatif jawaban. Disarankan jumlah alternatif jawaban minimal 1,5 kali dari jumlah pertanyaan stem soal.

- Jika jumlah butir soal cukup banyak, maka usahakan pertanyaan atau stem soal terletak dalam satu halaman dengan alternatif jawaban.

d. Tes Aebab Akibat

Tes sebab akibat merupakan tes yang menganalisis keterkaitan antara pernyataan pada satu bagian soal dengan pernyataan pada bagian yang lain. Soal ini terdiri dari stem soal dan alternatif jawaban. Stem soal umumnya disusun dalam 2 bagian, yakni bagian pertama yang menyatakan sebab, dan pernyataan kedua yang merupakan bagian akibat. Peserta tes kemudian diminta melakukan analisis terhadap nilai kebenaran dari masing-masing pernyataan pada stem soal, kemudian menilai keterkaitan antara pernyataan yang menjadi sebab dengan pernyataan yang menjadi akibat. Hasil analisis tersebut kemudian dinyatakan dalam bentuk alternatif jawaban. Pilihan jawaban menggambarkan keterkaitan antara pernyataan sebab dengan pernyataan akibat, yang biasanya diletakkan pada bagian petunjuk soal. Misalnya:

- Pilih A jika pernyataan pertama benar dan pernyataan kedua benar, serta keduanya memiliki hubungan sebab akibat.
- Pilih B jika pernyataan pertama benar dan pernyataan kedua benar tetapi keduanya tidak memiliki hubungan sebab akibat.
- Pilih C jika pernyataan pertama benar dan pernyataan kedua salah.
- Pilih D jika pernyataan pertama salah dan pernyataan kedua benar.
- Pilih E jika pernyataan kedua pernyataan salah.

Contoh soal:

Ibukota Negara RI adalah Jakarta

sebab

Proklamasi kemerdekaan RI dilakukan pada tanggal 17 Agustus 1945.

Dalam contoh di atas, kalimat "Ibukota Negara RI adalah Jakarta" adalah kalimat pernyataan sebab yang bernilai benar, dan kalimat "Proklamasi kemerdekaan RI dilakukan pada tanggal 17 Agustus 1945" merupakan pernyataan akibat yang bernilai benar, tetapi antara kedua pernyataan tersebut tidak berkaitan sebab akibat satu sama lain.

Beberapa kelebihan tes bentuk sebab akibat antara lain:

- Mampu menggambarkan kemampuan berfikir analitis dan asosiatif dari peserta tes
- Cocok digunakan untuk mengukur hasil belajar kognitif pada aspek yang kompleks
- Mudah mengoreksi dan memberi skor, karena jawaban peserta tes dapat dilihat langsung dari alternatif jawaban yang dipilih.

Sedangkan beberapa kelemahannya antara lain:

- Evaluator cukup sulit menyusun butir tes yang baik.
- Peserta tes cukup sulit menjawab atau mengerjakan butir tes.
- Masih memungkinkan peserta tes untuk menjawab dengan tebakan (*guessing).*
- Kurang cocok untuk butir tes yang menanyakan kemampuan berhitung.

Beberapa hal yang harus diperhatikan saat menyusun butir tes berbentuk sebab akibat antara lain:

- Susun pernyataan sesuai dengan tujuan pengukuran
- Pernyataan sebab dan akibat merupakan kalimat pendek yang jelas, tegas, dan hanya mengandung satu pokok masalah.
- Pernyataan sebab dan pernyataan akibat harus merupakan kalimat yang telah memiliki nilai kebenaran. Jangan menggunakan kata "jika-maka" sebagai kata penghubungnya.
- Segera buat kunci jawaban dan pedoman penskorannya pada saat membuat stem soal.

e. Tes Uraian

Tes uraian atau sering juga disebut essay adalah tes yang jawabannya membutuhkan jawaban dari peserta tes secara terurai dan rinci dengan kata-kata sendiri. Stem soalmya terdiri dari suatu pertanyaan atau pernyataan yang mebutuhkan jawaban rinci, panjang dan terurai.

Tes bentuk uraian lebih cocok digunakan untuk hasil belajar kognitif yang lebih kompleks, seperti aspek-aspek C4, C5 dan C6, terutama untuk mengukur kemampuan-kemampuan analitis. Tes uraian dapat dibedakan dengan tes isian. Tes uraian mebtuhkan jawaban panjang dan deskriptif, sedangkan tes isian hanya membutuhkan ingatan dan pemahaman karena merupakan penggalan suatu kalimat utuh.

Beberapa keunggulan tes uraian antara lain:

- Relatif mudah menyiapkan dan menyusun pertanyaannya.

- Mampu mengukur hasil belajar kognitif pada aspek yang kompleks.
- Membantu peserta tes untuk menumbuhkan kemampuan menulis dan berfikir kreatif.
- Memperkecil peluang menjawab secara tebakam

Sedangkan beberapa kelemahannya antara lain:

- Cakupan materi tes lebih terbatas.
- Jawaban yang panjang oleh peserta tes menyebabkan evaluator sulit menetapkan jumlah butir tes dalam jumlah banyak.
- Membutuhkan waktu relatif lama dalam menjawab soal tes.
- Meskipun jawaban pasti sudah ditetapkan oleh evaluator, tetapi variasi jawaban peserta tes dapat bersifat interpretatif sehingga bisa menyulitkan pemberian skor dan mempengaruhi obyektivitas evaluator.
- Relatif membutuhkan waktu dan energi untuk koreksi jawaban.
- Skor hasil pengukuran dapat terganggu oleh kualitas tulisan dan kemampuan menyusun kalimat dari peserta tes.
- Skor hasil pengukuran dapat dipengaruhi oleh kemampuan peserta tes dalam memahami pertanyaan secara tepat.
- Memungkinkan terjadinya *halo-effect,* yakni kecenderungan untuk memberi skor lebih tinggi pada peserta tes yang dikenal sebagai peserta tes yang pintar. Demikian juga sebaliknya pada peserta tes yang kurang pintar.
- Pemberian skor yang dapat bersifat subyektif menyebabkan tingkat ketepatan (validitas) dan

konsistensi (reliabilitas) hasil pengukurannya rendah.

Beberapa kaidah yang harus diperhatikan dalam menyusun butir tes berbentuk uraian adalah:

- Susun butir tes sesuai dengan tujuan pengukuran.
- Pastikan bahwa tbentuk tes uraian cocok digunakan sesuai dengan karakteristik peserta tes dan materi yang diajarkan.
- Gunakan kalimat pertanyaan atau perintah yang singkat, jelas, dan hanya mengandung satu pokok masalah sehingga tidak menimbulkan interpretasi ganda. Hindari penggunaan kalimat atau kata "Apa yang anda ketahui", "coba Anda jelaskan..", dan semacamnya, karena kata atau kalimat tersebut mengandung ketidakpastian.
- Hindari melakukan *copy-paste* kalimat dari buku atau teks yang telah ada, karena dapat meningkatkan peluang peserta tes untuk menyontek atau menebak jawaban yang dianggapnya paling tepat.
- Menyusun kunci jawaban dan pedoman penskorannya segera setelah butir soal dibuat.
- Jangan memberikan kesempatan kepada peserta tes untuk memilih beberapa butir tes untuk dijawab, karena hal itu menyebabkan bias pengukuran.

2. Bentuk-Bentuk Non Tes

Instrumen non tes umumnya digunakan untuk mengukur hasil belajar aspek afektif, karena hasil pengukurannya yang bersifat selalu benar. Artinya, skor dan skala hasil pengukuran yang dikembangkan, tidak menunjukkan bahwa seseorang telah mencapai kompetensi atau hasil belajar tertentu, tetapi lebih menunjukkan *grade* atau kategorinya pada aspek afektif

yang diukur pada saat itu. Dengan kata lain, hasil pengukuran aspek-aspek afektif menunjukkan kecenderungan afeksi seseorang pada suatu waktu.

Instrumen non tes yang umum digunakan berbentuk:

a. Angket atau Kuisioner

Angket dapat dipandang sebagai teknik pengukuran maupun sebagai instrumen. sebagai teknik pengukuran, angket dapat diartikan sebagai cara yang digunakan evaluator untuk mengukur aspek-aspek psikologis, terutama aspek afektif, menggunakan sejumlah pertanyaan tertulis. Dalam artian instrumen, angket atau kuisioner adalah sejumlah pertanyaan atau pernyataan tertulis yang diberikan kepada sekelompok orang (sering disebut responden) untuk dijawab atau direspon secara tertulis. Dengan kata lain, angket pada dasarnya adalah wawancara tertulis.

Sebagaimana butir soal tes, angket terdiri dari 2 bagian, yakni bagian batang tubuh (stem) dan bagian jawaban.

Ditinjau dari cara menjawabnya, angket dapat dibedakan menjadi angket tertutup dan angket terbuka. Kadangkala evaluator dapat menggunakan kombinasi antara keduanya, untuk memperoleh gambaran yang lebih jelas tentang aspek yang diukur. Angket kombinasi tersebut sering disebut sebagai angket berbentuk semi tertutup. Dalam buku ini hanya akan dibahas tentang angket tertutup dan angket terbuka, yakni sebagai berikut:

Angket Tertutup

Angket tertutup adalah angket yang jawabannya telah disediakan, sehingga responden tinggal memilih atau memberi tanda pada salah satu atau beberapa alternatif jawaban yang disediakan. Penggunaan angket tertutup ditujukan agar hasil belajar afektif bisa dinyatakan dalam bentuk skor. Selanjutnya skor tersebut kemudian dapat dinyatakan menjadi nilai dalam bentuk kategori dan klasifikasi. Misalkan kita ingin mengetahui kategori sikap seorang peserta didik terhadap nata pelajaran IPA, maka dengan menggunakan angket tertutup kita akan memperoleh sebaran skor sikap peserta didik, sehingga pada akhirnya dapat disimpulkan bahwa sikap seorang peserta didik adalah cenderung positif atau negatif.

Kelebihan angket tertutup antara lain:

- Praktis dalam penggunaannya
- Tidak memerlukan kehadiran evaluator.
- Bisa melakukan pengukuran dalam jumlah banyak secara serentak dan bersamaan.
- Responden memiliki kebebasan waktu dan keadaan untuk mengisi angket.
- Mudah dalam memberikan skor dan menilai
- Penilaian lebih obyektif
- Responden cenderung lebih mudah menjawab
- Dapat dibuat secara online

Adapun beberapa kekurangannya antara lain:

- Responden tidak memiliki kebebasan untuk mengemukakan pendapat dan alasan.

- Sulit mendapatkan jaminan bahwa jawaban yang diberikan adalah jujur dan sesuai dengan keadaan atau karakteristik responden yang diukur.
- Kadangkala tidak semua responden bersedia mengisi dan mengembalikan angket kepada evaluator.
- Tergantung dengan karakristik responden, misalnya kemampuan membaca dan memahami makna kalimat pertanyaan/pernyataan.
- Terdapat peluang fakabilitas, yakni adanya kecenderungan responden menjawab sesuatu yang menyenangkan evaluator

Beberapa kaidah yang harus diperhatikan dalam penyusunan angket tertutup antara lain:

- Susunlah pertanyaan/pernyataan angket sesuai dengan tujuan pengukuran dan karakteristik responden.
- Berikan petunjuk pengisian angket secara jelas pada bagian awal.
- Susunlah alternatif jawaban yang sesuai dengan stem pertanyaan/pernyataan.
- Gunakan kata atau kalimat yang tegas, jelas, ringkas, dan hanya mengandung satu pokok masalah sehingga mudah difahami responden.
- Hindari penggunaan kata atau kalimat sugestif atau memojokkan responden.

Angket Terbuka

Angket terbuka adalah adalah angket yang jawabannya belum disediakan, sehingga responden secara bebas menyusun kalimat jawaban sesuai dengan yang mereka inginkan. Angket bentuk ini terdiri dari bagian stem berupa sejumlah pertanyaan

atau pernyataan, dan bagian jawaban angket yang umumnya berupa tempat kosong atau titik-titik.

Angket terbuka digunakan untuk memperoleh jawaban deskriptif dan mendalam dari responden, misalnya tentang pendapat, pandangan, dan kritik terhadap sesuatu. Dalam konteks pengukuran sikap, angket terbuka digunakan untuk mendapatkan gambaran tentang latar belakang sikap yang dimiliki responden, atau mengapa responden bersikap demikian. Dengan demikian, angket bentuk ini tidak dimaksudkan untuk diberikan skor tertentu, tetapi untuk mendapatkan gambaran mendalam tentang sikap tersebut. Dalam konteks tersebut, kita tidak menggunakan istilah pengukuran tetapi menggunakan istilah evaluasi.

Beberapa kelebihan angket terbuka antara lain:

- Relatif mudah menyusun pertanyaan atau pernyataannya.
- Dapat mengukur aspek-aspek afeksi secara mendalam.
- Responden memiliki kebebasan waktu dan keadaan untuk mengisi angket.
- Responden lebih bebas mengungkapkan pendapat dan kritik, serta mendeskripsikan keadaannya.

Adapun beberapa kelemahannya antara lain:

- Jumlah pertanyaan angket terbatas.
- Relatif sulit memberikan skor terhadap jawaban responden.
- Proses koreksi dan analisis jawaban responden cenderung memerlukan waktu dan tenaga.
- Hasil analisis terhadap jawaban responden cenderung subyektif dan dipengaruhi oleh faktor

lain seperti kerapian tulisan, kemampuan responden menyusun kalimat, dan sebagainya.

- Dapat menyebabkan responden kesulitan menjawab dan menyusun kalimat jawaban.
- Responden dapat bosan menjawab pertanyaan dalam jumlah banyak, sehingga akan cenderung sembarangan menjawab pertanyaan-pertanyaan selanjutnya.

Beberapa kaidah yang harus diperhatikan dalam menyusun angket terbuka antara lain:

- Susunlah butir angket sesuai dengan tujuan evauluasi dan karakteristik responden.
- Berikan petunjuk pengisian secara jelas pada bagian awal angket.
- Jumlah butir angket tidak terlalu banyak.
- Susunlah kalimat stem dalam bentuk pertanyaan/pernyataan yang singkat, jelas, tegas dan hanya mengandung satu pokok masalah, sehingga tidak menimbulkan interpretasi ganda.
- Susunlah *frame* atau kerangka jawaban dari pertanyaan angket sebagai patokan.

b. Wawancara dan Pedoman Wawancara

Wawancara merupakan teknik pengukuran, yang dapat diartikan sebagai cara yang digunakan evaluator untuk mengukur aspek-aspek psikologis, terutama aspek afektif, menggunakan sejumlah pertanyaan lisan yang dijawab juga secara lisan oleh responden. Hasil wawancara tersebut dapat dinyatakan dalam bentuk skor maupun secara deskriptif. Sedangkan sebagai instrumen wawancara, digunakan pedoman wawancara, yang diartikan sebagai panduan tertulis yang berisi sejumlah butir pertanyaan/pernyataan yang akan

ditanyakan kepada responden melalui wawancara. Dengan demikian, ada perbedaan mendasar antara wawancara dengan pedoman wawancara. Wawancara merupakan teknik pengukuran, sedangkan pedoman wawancara merupakan instrumen pengukuran.

Ditinjau dari teknik penggunaan dan substansinya, wawancara dapat dibedakan menjadi wawancara terstruktur dan wawancara mendalam.

Wawancara Terstruktur

Wawancara terstruktur adalah wawancara yang mana jawaban responden cenderung diarahkan pada jawaban tertentu. Dalam wawancara terstruktur, responden diarahkan dan diberi pilihan untuk menjawab dengan jawaban tertentu atau yang sudah disiapkan. Contoh:

Pertanyaan lisan	Jawaban lisan responden
Apakah Anda selalu sarapan pagi ?	Selalu Sering Kadang-kadang Jarang Tidak pernah
Bagaimana pendapat Anda tentang pembelajaran online?	Sangat baik Baik Cukup baik Kurang baik

Dari contoh di atas, sebenarnya wawancara terstruktur memiliki kemiripan dengan angket tertutup. Perbedaannya hanya pada penggunaannya, karena wawancara digunakan secara lisan dan dijawab secara lisan.

Wawancara terstruktur umumnya digunakan jika evaluator ingin memberikan skor tertentu terhadap hasil wawancara. Dengan menggunakan pola pemberian skor bergradasi, sebagaimana juga pemberian skor pada angket tertutup, maka keseluruhan jawaban responden dapat diberikan skor tertentu.

Beberapa kelebihan wawancara terstruktur adalah:

- Praktis penggunaannya, terutama jika jumlah responden tidak terlalu banyak dan memungkinkan untuk diwawancarai.
- Mudah dalam memberikan skor dan menilai
- Pemberian skor dan penilaian lebih obyektif
- Responden cenderung lebih mudah menjawab secara lisan karena telah disediakan pilihan jawaban.

Adapun beberapa kekurangannya antara lain:

- Responden tidak memiliki kebebasan untuk mengemukakan pendapat dan alasan.
- Teknik wawancara dengan komunikasi langsung dapat menyebabkan kendala psikologis seperti gugup dan kecemasan responden, sehingga mempengaruhi hasil pengukuran.
- Terdapat peluang fakabilitas, yakni adanya kecenderungan responden menjawab sesuatu yang menyenangkan evaluator.

Beberapa kaidah yang harus diperhatikan dalam penggunaan wawancara terstruktur antara lain:

- Susunlah pedoman wawancara sebagai instrumen pemandu.

- Pertanyaan/pernyataan dalam panduan wawancara harus sesuai dengan tujuan pengukuran dan karakteristik responden.
- Berikan pengantar dan penjelasan secara jelas pada bagian awal pedoman wawancara. Sampaikan petunjuk tersebut kepada responden sebelum pelaksanaan wawancara.
- Susunlah alternatif jawaban yang sesuai dengan stem pertanyaan/pernyataan.
- Gunakan kata atau kalimat yang tegas, jelas, ringkas dan mudah difahami responden.

Wawancara Mendalam (*indeepth interview*).
Wawancara mendalam merupakan wawancara yang mana responden diberi kebebasan untuk menjawab setiap pertanyaan. Pertanyaan atau pernyataan yang diberikan dalam wawancara mendalam bersifat terbuka sehingga memungkinkan responden menjawab secara luas, rinci, dan mendalam. Pertanyaan tersebut dapat bersifat memancing dan menstimulasi jawaban responden. Pertanyaan-pertanyaan dalam wawancara mendalam juga dapat berkembang berdasarkan jawaban terakhir responden, sehingga pedoman wawancara hanya bersifat pertanyaan stumulus awal.

Contoh:

Pertanyaan lisan evaluator	Jawaban lisan responden
Pertanyaan utama: Bagaimana pendapat Anda tentang model pembelajaran online yang selama ini dilaksanakan? **Pertanyaan lanjutan:** Mengapa Anda berpendapat demikian? Menurut Anda, faktor apa saja yang	……

harus diperhatikan dalam pembelajaran online? Faktor apa yang mendukung pembelajaran online? Faktor apa yang menghambat pembelajaran online? Fasilitas apa yang harus disiapkan agar pembelajaran online tersebut efektif?	

Dari contoh di atas, tampak bahwa pertanyaan utama sebenarnya merupakan pertanyaan stimulus awal, yang kemudian diteruskan atau didalami dengan pertanyaan-pertanyaan lanjutan. Itulah sebabnya wawancara model ini disebut sebagai wawancara mendalam, karena adanya pertanyaan-pertanyaan lanjutan yang digunakan sebagai pendalaman.

Wawancara mendalam umumnya digunakan untuk mendapatkan gambaran utuh tentang hasil belajar aspek afektif. Hasil wawancara tidak dinyatakan dalam bentuk skor hasil pengukuran, tetapi dinyatakan dalam bentuk deskriptif. Hasil wawancara dianalisis secara mendalam dan analitik, kemudian dipaparkan secara kualitatif untuk menggambarkan afeksi responden.

Beberapa kelebihan penggunaan wawancara mendalam sebagai teknik evaluasi adalah:

- Wawancara yang dilaksanakan secara langsung dari hati ke hati, bisa membuat responden lebih bebas mengemukakan pendapat dan perasaannya.
- Dapat menggambarkan aspek afektif beserta faktor-faktor terkait secara mendalam

- Dapat menggambarkan karakteristik afeksi responden secara lebih komprehensif.

Sedangkan beberapa kelemahannya antara lain:

- Sulit digunakan jika responden berjumlah banyak.
- Relatif sulit menyusun pertanyaan/pernyataan, khususnya pertanyaan lanjutan.
- Penilaian bersifat kualitatif sehingga evaluator sulit menjaga obyektivitas. Masalah ini dapat diatasi dengan melakukan cross-chek antar satu jawaban dengan jawaban yang lain.
- Penilaian dapat dipengaruhi keterampilan responden, antara lain keterampilan bahasa lisan dan menyusun kalimat.

Beberapa kaidah yang harus diperhatikan dalam penggunaan wawancara mendalam antara lain:

- Susunlah pedoman wawancara sebagai instrumen pemandu.
- Susunlah butir-butir wawancara sesuai dengan tujuan evaluasi dan karakteristik responden.
- Hindari menyusun pertanyaan lanjutan yang terlalu banyak.
- Pertanyaan/pernyataan dalam panduan wawancara harus sesuai dengan tujuan pengukuran dan karakteristik responden.
- Berikan pengantar dan penjelasan secara jelas pada bagian awal pedoman wawancara. Sampaikan petunjuk tersebut kepada responden sebelum pelaksanaan wawancara.
- Gunakan kata atau kalimat yang tegas, jelas, ringkas dan mudah difahami responden.

- Hindari penggunaan kata atau kalimat yang bersifat sangat pribadi dan memojokkan responden.

c. Dokumentasi dan Pedoman Dokumentasi

Dokumentasi dapat dipandang sebagai teknik evaluasi, yakni suatu cara pengumpulan data dari catatan, buku, transkrip, foto, berkas, dan dokumen-dokumen lainnya. Sedangkan sebagai panduan dokumentasi, maka evaluator terlebih dahulu menyusun pedoman dokumentasi. Dengan kata lain, dokumentasi merupakan teknik evaluasi, sedangkan sebagai instrumennya adalah pedoman dokumentasi. Penggunaan pedoman dokumentasi sebagai instrumen dimaksudkan agar evaluasi atau pengumpulan data lebih terarah sesuai dengan tujuan pengukuran. Pada kenyataannya di lapangan, dokumentasi dapat berkembang sesuai kebutuhan, sehingga pedoman dokumentasi lebih bersifat sebagai panduan awal. Prinsip-prinsip dokumentasi untuk dapat memotret keadaan lebih lengkap, lebih rinci, dan lebih banyak, menyebabkan pedoman dokumentasi dapat dikembangkan sesuai kebutuhan keadaan saat evaluasi dilakukan.

Kelebihan metode dokumentasi sebagai teknik evaluasi antara lain:

- Dapat digunakan untuk menggambarkan keadaan responden pada masa lampau (bersifat longitudinal), sehingga karakteristik responden dapat dipetakan secara lebih lengkap.
- Relatif lebih obyektif menggambarkan keadaan masa lampau karena tidak berhubungan langsung dengan responden. Pada saat dokumentasi, responden tidak merasa bahwa dia sedang diamati atau diukur.

- Cukup mudah digunakan karena bahan dokumentasi sudah tersedia dan beragam.

Sedangkan kelemahannya antara lain:

- Catatan dan dokumen yang diteliti perlu dipilah secara hati-hati oleh evaluator, karena tidak terdapat keyakinan yang kuat tentang akurasi pencatatan yang telah dilakukan orang lain di masa lampau. Hal ini terjadi karena tujuan pengukuran yang dilakukan evaluator saat ini, tidak selalu sama dengan tujuan orang lain melakukan dokumentasi pada masa lalu.
- Dokumen yang ditelaah masih bersifat interpretatif, sehingga membutuhkan kejelian evaluator dalam menganalisis dan menginterpretasi dokumen tersebut. Misalkan selembar foto, bisa menggambarkan berbagai situasi masa lampau.
- Sulit untuk pemberian skor dan melakukan penilaian.

Kelebihan dokumentasi sebagai teknik pengukuran hasil belajar antara lain:

Adapun kaidah yang harus diperhatikan dalam penggunaan teknik dokumentasi dan menyusun panduan dokumentasi, antara lain:

- Susun pedoman dokumentasi sesuai dengan tujuan evaluasi dan karakteristik responden.
- Penyusunan pedoman dokumentasi dijadikan sebagai panduan awal dan dapat dikembangkan sesuai dengan kebutuhan pengukuran.
- Buat daftar *checklist* yang menggambarkan dokumen mana yang telah dan belum terkumpul.

- Mengingat banyaknya dokumen yang tersedia, maka kumpulkan hanya dokumen yang benar-benar relevan dengan tujuan evaluasi.

d. Observasi dan Pedoman Observasi

Observasi adalah suatu teknik evaluasi berdasarkan pengamatan evaluator terhadap obyek pengukuran dan lingkungan sekitarnya. Pengamatan tersebut dapat berupa melihat, mencatat, merekam, dan mengukur kejadian di lapangan. Observasi dapat dilakukan terhadap orang, benda, dan keadaan.

Observasi dapat dibedakan menjadi observasi partisipan dan observasi non partisipan.

- Observasi partisipan adalah observasi dimana evaluator atau observer menjadi bagian dari keadaan atau kehidupan responden yang diteliti. Hubungan antara pengamat dengan responden adalah saling terkait, karena keduanya berada pada keadaan yang sama. Jenis observasi ini umumnya digunakan dalam penelitian kualitatif seperti penelitian etnografi. Misalkan penelitian tentang kearifan lokal pada masyarakat suku Dayak di Kalimantan, maka dalam hal ini observer harus membaur ke dalam kehidupan responden yang diteliti sehingga keaslian data akan muncul dan didapatkan.
- Observasi non partisipan adalah observasi yang memisahkan antara observer dengan responden. Artinya dalam hal ini posisi observer adalah sebagai pengamat di luar kejadian yang diamati, hubungan antara pengamat dan responden adalah saling bebas. Jenis observasi ini bisa digunakan dalam banyak kajian dan evaluasi, termasuk dalam evaluasi pembelajaran. Misalkan seorang guru mengamati penerapan perilaku sopan santun pada peserta didiknya.

Untuk melakukan observasi, dibutuhkan pedoman observasi sebagai instrumen. pedoman observasi memuat seperangkat pertanyaan atau pernyataan perintah tentang apa yang harus diamati evaluator. Dalam melaksanakan observasi terhadap benda hidup, evaluator harus mengupayakan agar orang yang diamati tidak merasa sedang diamati, sehingga dapat dijamin keaslian dan kewajaran data yang terkumpul. Untuk itu, kadangkala evaluator membutuhkan alat bantu tertentu, seperti kamera CCTV, untuk merekam keadaan yang asli.

Kelebihan observasi antara lain adalah:

- Dapat mengamati keadaan secara langsung sehingga lebih obyektif, terutama jika responden tidak mengetahui bahwa sedang diobservasi.
- Pencatatan hasil pengamatan dapat dilakukan secara langsung sehingga mengurangi risiko pembiasan.
- Dapat digunakan secara bebas tanpa mengganggu waktu dan tenaga responden.
- Jika menggunakan peralatan seperti kamera, video, atau alat perekam lainnya, maka pengamatan dapat dilaksanakan secara serentak di berbagai tempat.

Sedangkan kelemahan observasi antara lain adalah:

- Memungkinkan terjadinya fakabilitas, khususnya jika responden mengetahui bahwa ia sedang diobservasi.
- Hanya cocok untuk kejadian yang berjangka waktu relatif pendek. Untuk kejadian yang berlangsung lama, observasi kurang memungkinak untuk merekam seluruh kejadian secara lengkap. Untuk kejadian yang berjangka

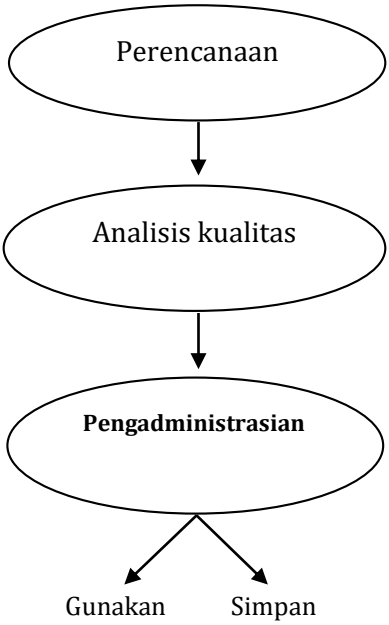
panjang, harus digunakan teknik observasi partisipan.

- Pengamatan dapat terganggu jika terjadi peristiwa yang tidak terduga, baik terhadap keadaan, responden, maupun kejadian tiba-tiba yang menimpa evaluator.

D. Langkah-langkah Penyusunan Instrumen Pengukuran

Penyusunan instrumen pengukuran dan evaluasi, baik instrumen tes maupun non tes, harus dilakukan dengan tahapan yang benar. Penyusunan secara bertahap dan benar dapat menjamin kualitas instrumen tersebut.

Secara garis besar, penyusunan instrumen pengukuran terbagi menjadi beberapa langkah sebagaimana diagram berikut:



Gambar 4.1. Diagram langkah penyusunan instrumen

Langkah-langkah penyusunan instrumen adalah sebagai berikut:

1. Perencanaan

a. Menetapkan tujuan pengukuran

Tahap penetapan tujuan pengukuran merupakan tahapan yang paling penting dalam perencanaan instrumen, karena akan mengarahkan kegiatan pengukuran atau evaluasi kepada hal-hal teknis yang harus ditempuh. Tujuan pengukuran menggambarkan kondisi hasil pembelajaran yang diharapkan dan yang akan kita ukur. Dengan demikian, kata kunci tujuan pengukuran adalah adanya kesesuaiannya dengan tujuan instruksional dan kompetensi hasil belajar yang diharapkan. Artinya, penyusunan tujuan pengukuran harus disesuaikan dengan tujuan pembelajaran dan kompetensi yang ingin dicapai sebagai hasil pembelajaran. Tujuan pengukuran umumnya juga merupakan pernyataan operasional dari tujuan pembelajaran. Jika diperlukan, pada tujuan pengukuran dapat ditambahkan deskripsi indikator yang akan diukur.

Contoh:

Tujuan pembelajaran	Tujuan pengukuran
Peserta didik mampu menyebutkan nama ibukota provinsi Banten.	Mengetahui kemampuan peserta didik untuk menyebutkan nama ibukota provinsi Banten.
Peserta didik mampu menjumlahkan pecahan biasa berpenyebut sama	Mengetahui kemampuan peserta didik untuk menjumlahkan pecahan biasa dan berpenyebut sama
Peserta didik memiliki sikap jujur	Mengetahui adanya sikap jujur peserta didik

	dengan indikator: - Selalu jujur dalam perkataan - Selalu jujur dalam tindakan - Selalu obyektif dalam menilai sesuatu.
Peserta didik memiliki sikap toleransi	Mengetahui adanya sikap toleransi pada peserta didik, dengan indikator: - Menghargai perbedaan - Menghargai pendapat oranglain - Tidak memaksakan kehendak - Bersedia menerima kritik
Peserta didik memiliki keterampilan mengetik dengan kecepatan minimal 200 kata per menit menggunakan komputer.	Mengukur keterampilan mengetik peserta didik dengan ukuran: - 200 kata permenit - Menggunakan komputer - Menggunakan program MS-Word

Pada contoh di atas, beberapa tujuan pengukuran dijabarkan dengan menambah indikator-indikator hasil belajar yang akan diukur. Penambahan ini penting dilakukan, terutama jika evaluator menganggap bahwa tujuan pembelajaran atau kompetensi hasil belajar yang dinyatakan, masih belum operasional. Dengan

demikian, tujuan pengukuran lebih bersifat sebagai operasionalisasi dari tujuan pembelajaran.

Kata kunci dari penyusunan tujuan pengukuran yang operasional adalah penggunaan kata atau kalimat operasional yang tepat untuk masing-masing aspek yang diukur, sebagaimana telah dipaparkan pada bab sebelumnya tentag aspek-aspek pengukuran.

Dalam beberapa keadaan, evaluator tidak merumuskan tujuan pengukuran secara eksplisit. Pada sebagian evaluator yang sudah berpengalaman, tujuan pengukuran telah dirumuskan secara otomatis dari tujuan pembelajaran sehingga sudah berada dalam otak dan fikirannya. Tetapi bagi evaluator pemula, sangat disarankan untuk menyusun tujuan pengukuran dalam perencanaan instrumen, sehingga dapat menjamin kualitas instrumen pengukuran.

2. Menentukan Aspek Hasil Belajar yang Akan Diukur

Sebagaimana telah dijelaskan sebelumnya, aspek hasil belajar peserta didik terdiri dari aspek pengetahuan (kognitif), sikap (afektif), dan keterampilan (psikomotorik). Aspek kognitif dapat diklasifikasi menjadi 6 sub aspek atau tahapan berfikir, yakni C1 (mengingat), C2 (memahami), C3 (mengaplikasikan), C4 (menganalisis), C5 (mengeveluasi), dan C6 (mencipta). Sedangkan aspek sikap, diklasifikasikan menjadi 5 sub aspek atau tahapan bersikap, yakni A1 (menerima), A2 (menanggapi), A3 (menghargai), A4 (mengorganisir), dan A5 (karakterisasi). Adapun aspek keterampilan diklasifikasi menjadi 4 sub aspek atau tahapan berperilaku, yakni P1 (menirukan), P2 (memanipulasi), P3 (mengalamiahkan), dan P4 (mengartikulasikan).

Mengingat luasnya aspek-aspek pengukuran beserta tahapan-tahapannya, maka evaluator terlebih dahulu menetapkan aspek dan tahapan mana yang akan diukur. Tentu saja penentuan aspek dan tahapan ini disesuaikan

dengan tujuan instruksional atau standar kompetensi yang telah ditetapkan sebagai tujuan pembelajaran.

Penentuan aspek dan tahapan hasil belajar yang akan diukur, akan berpengaruh pada keputusan tentang hal-hal tekni pengukuran. Misalkan untuk menentukan jenis instrumen yang akan digunakan, maka evaluator harus menyesuaikannya dengan aspek yang akan diukur. Hal itu digambarkan sebagai berikut:

Aspek yang akan diukur		Jenis instrumen
Kognitif	→	Tes
Afektif	→	Non tes
Keterampilan	→	Tes kinerja

Selanjutnya evaluator dituntut untuk menentukan bentuk instrumen yang cocok untuk digunakan. Faktor yang dipertimbangkan dalam menentukan bentuk instrumen adalah:

- Karakteristik peserta tes atau responden. Misalnya, bentuk tes uraian uraian tidak cocok digunakan pada peserta tes kelas II SD, karena kemampuan membaca dan menalar mereka relatif masih terbatas. Angket terbuka yang membutuhkan jawaban deksriptif-analitik, tentu juga tidak cocok digunakan pada responden murid SD. Untuk mengukur sikap murid SD, lebih cocok digunakan wawancara, terutama jika respondennya tidak terlalu banyak.
- Tahapan hasil belajar yang akan diukur. Evaluator perlu menentukan tahapan atau sub aspek hasil belajar yang akan diukur, karena hal ini akan menentukan bentuk instrumen yang cocok untuk digunakan. Misalkan evaluator akan menggunakan

instrumen jenis tes untuk mengukur aspek kognitif, sedangkan sub aspek yang akan diukur adalah sub aspek C4 (analisis). Sub aspek C4 sebaiknya diukur menggunakan tes berbentuk uraian. Tes uraian memiliki kemampuan untuk mengukur kemampuan analitik.

- Jumlah peserta tes atau responden. Berdasarkan pertimbangan teknis, instrumen bentuk tertentu sulit digunakan jika jumlah peserta tes atau respondennya cukup banyak. Misalnya tes bentuk uraian, sulit digunakan untuk peserta tes yang jumlahnya banyak, karena kesulitan evaluator dalam pemberian skor dan penilaian. Kesulitan semacam ini kadangkala diatasi dengan cara membagi peserta tes ke dalam kelompok-kelompok kecil, sehingga secara teknis, tes tersebut lebih mudah dilaksanakan dan hasilnya lebih mudah diberi skor atau dinilai. Dapat pula evaluator menunjuk beberapa orang *rater* atau pengkoreksi, sehingga beban untuk melakukan penilaian terhadap jawaban peserta tes dapat dibagi dan terkurangi. Hal yang sama juga mungkin terjadi pada saat evaluator akan menggunakan angket terbuka kepada responden yang jumlahnya banyak.

Dalam kaitannya dengan pengukuran hasil belajar koginitif, keterkaitan antara sub aspek pengukuran dan bentuk tes dapat digambarkan sebagai berikut:

Sub aspek yang diukur		Bentuk tes yang digunakan
C1 (mengingat)	→	PG, isian, menjodohkan
C2 (memahami)	→	PG, isian, menjodohkan
C3 (mengapilkasikan)	→	PG, isian, menjodohkan,

		uraian singkat
C4 (menganalisis)		Uraian, sebab akibat
C5 (mengevaluasi)		Uraian, sebab akibat
C6 (mencipta)		Uraian, sebab akibat

Tentu saja hubungan di atas merupakan sebuah panduan, tetapi bukan sebuah aturan. Artinya, faktor sub aspek yang akan diukur bukan satu-satunya faktor yang dipertimbangkan dalam penggunaan bentuk tes tertentu, sehingga hubungan di atas dapat disesuaikan dengan kebutuhan dan keadaan.

3. Menentukan Cakupan/Kedalaman Materi yang Akan Diukur

Dalam merencanakan dan menyusun instrumen pengukuran, evaluator juga harus menentukan cakupan luas dan kedalaman materi atau substansi yang akan diukur. Kadangkala, dengan adanya berbagai keterbatasan teknis, misalnya jumlah waktu yang tersedia, maka evaluator tidak dapat menyusun instrumen yang dapat mencakup keseluruhan isi materi yang hendak diukur. Dengan demikian, evaluator harus memilih sebagian materi yang dianggap penting atau dapat mewakili keseluruhan materi ukur.

Pemilihan cakupan luas dan kedalaman materi yang akan diukur, berpengaruh pada jenis, bentuk, dan jumlah butir instrumen yang akan digunakan. Misalkan dalam pengukuran hasil belajar aspek kognitif, kelebihan dan kekurangan tes berbentuk pilihan ganda dan isian, dapat dibandingkan dengan kelebihan dan kekurangan tes uraian dan sebab akibat:

Cakupan materi ukur		Bentuk tes yang digunakan
Luas	→	PG, isian singkat
Luas dan dangkal	→	PG, isian singkat
Terbatas/sempit	→	Uraian, sebab akibat
Dalam	→	Uraian, sebab akibat

Dengan demikian, butir-butir soal PG dan isian umumnya dapat digunakan untuk mengukur cakupan materi evaluasi yang luas. Soal tes berbentuk PG dan isian cenderung mudah dan cepat mengerjakannya, sehingga evaluator dapat menyusun jumlah butir soal tes yang lebih banyak. Butir soal tes yang lebih banyak cenderung mampu mencakup materi ukur yang lebih luas. Sedangkan untuk menjawab soal tes berbentuk uraian, peserta tes membutuhkan waktu dan keterampilan menulis yang memadai. Akibatnya butir soal tes yang dapat disusun juga terbatas jumlahnya, sehingga cakupan materi ukur tes tersbeut juga terbatas. Tetapi, bentuk tes uraian dapat mengukur aspek-aspek hasil belajar yang lebih mendalam dan kompleks.

4. Menetapkan Jenis dan Bentuk Instrumen

Pada tahapan ini, evaluator harus menentukan jenis dan bentuk instrumen yang akan digunakan. Sebagaimana telah dijelaskan sebelumnya, jenis dan bentuk instrumen pengukuran dipengaruhi oleh beberapa faktor. Pemilihan jenis dan bentuk instrumen tersebut harus mempertimbangkan faktor antara lain:

- Aspek hasil belajar yang akan diukur. Apakah aspek yang diukur tersebut merupakan aspek kognitif, afektif, atau psikomotor?. Pengukuran aspek kognitif

dan keterampilan umumnya menggunakan tes, sedangkan pengukuran aspek afektif cenderung menggunakan non tes.

- Sub aspek hasil belajar yang akan diukur. Pada aspek kognitif misalnya, apakah instrumen yang disusun tersebut akan mengukur salah satu atau beberapa sub aspek C1, C2, C3, C4, C5, dan C6?.
- Tujuan akhir evaluasi. Perlu dikaji terlebih dahulu, apakah hasil evaluasi nantinya akan dinyatakan dalam secara kuantitatif ataukah secara kualitatif?. Jika hasil pengukuran nantinya akan dinyatakan dalam bentuk skor atau angka, maka instrumen yang digunakan harus menggunakan jawaban yang pasti, baik dalam bentuk alternatif jawaban yang telah disiapkan, maupun dalam bentuk jawaban singkat yang mudah diberi skor secara obyektif. Pada pengukuran aspek kognitif, bentuk soal PG dan isian singkat merupakan instrumen yang mudah untuk diberi skor. Pada pengukuran afektif, maka biasanya digunakan angket tertutup dengan model skala bergradasi atau skala sikap, yang mana setiap alternatif jawabannya merujuk pada skor tertentu.

Sebaliknya, jika evaluasi dimaksudkan untuk memberikan gambaran deksriptif tentang peserta tes atau responden, maka instrumen yang digunakan harus memiliki jawaban terbuka, deskriptif dan panjang. Jawaban deskriptif lebih mudah untuk dideskripsikan. Pada pengukuran aspek kognitif, soal-soal berbentuk uraian cenderung bisa mendeskripsikan jalan fikiran peserta tes. Pada pengukuran aspek afektif, angket terbuka maupun wawancara mendalam cenderung mampu mendeskripsikan afeksi secara lebih komprehensif dan mendalam. Pada pengukuran keterampilan, tes perbuatan yang bersifat terbuka (tanpa diarahkan langkah-langkahnya), merupakan instrumen

yang cocok digunakan untuk mengukur perilaku lebih mendalam.

- Jumlah peserta tes atau responden. Atas alasan teknis, jenis dan bentuk instrumen tertentu kurang praktis untuk digunakan pada peserta tes atau responden dalam jumlah banyak. Pada pengukuran aspek afektif misalnya, jika jumlah respondennya cukup banyak, maka teknik angket tertutup cenderung lebih praktis digunakan dibanding teknik wawancara maupun angket terbuka. Contoh lainnya adalah, tes berbentuk uraian lebih tepat jika digunakan pada peserta tes yang jumlahnya sedikit.

Meskipun demikian, dalam evaluasi kadangkala dibutuhkan campuran jenis dan bentuk instrumen. Misalnya dalam pengukuran hasil belajar kognitif sub aspek C2 dan C4, evaluator dapat saja menggunakan soal-soal bentuk PG ditambah dengan beberapa soal bentuk uraian. Kombinasi juga dapat dilakukan jika evaluator ingin mengukur aspek kognitif sekaligus aspek afektif, dengan cara menggunakan tes dan angket secara bersamaan.

5. Menetapkan Jumlah Butir

Jumlah butir pertanyaan atau pernyataan dalam instrumen, merupakan faktor yang perlu dipertimbangkan dalam penyusunan instrumen. Pertimbangan tersebut terkait dengan beberapa hal teknis berikut ini:

- Cakupan luas materi ukur. Semakin banyak butir pertanyaan/pernyataan, maka semakin luas materi ukur yang tercakup.
- Waktu yang tersedia dalam sekali pengukuran. Artinya, semakin banyak jumlah waktu yang tersedia untuk pelaksanaan evaluasi, maka memungkinkan untuk menggunakan jumlah butir yang lebih banyak. Dalam praktiknya, bisa saja waktu yang tersedia untuk

evaluasi cukup banyak tersedia, tetapi terbagi dalam beberapa tahapan waktu berbeda, misalkan evaluasi dilaksanakan setiap hari selama 3 hari, setiap harinya disiapkan waktu sebanyak 100 menit. Dengan jumlah waktu tersedia cukup banyak, maka evaluator dapat menyiapkan beberapa jenis dan bentuk instrumen berbeda pada setiap pengukuran.

- Jenis dan bentuk instrumen. Pada instrumen dengan bentuk jawaban terbuka (seperti tes uraian, angket terbuka, dan wawancara mendalam), jumlah butir pertanyaan/pernyataan sebaiknya dibatasi untuk menghindari kelelahan dan kebosanan responden dalam menjawab pertanyaan. Kebosanan dan kelelahan akan menurunkan akurasi jawaban. Selain itu, pertanyaan terbuka yang terlalu banyak juga akan menyebabkan kesulitan evaluator dalam menganalisis jawaban responden.

Dengan demikian, panduan tentang jumlah butir yang dapat digunakan sebagai patokan adalah sebagai berikut:

Jenis jawaban instrumen		Jumlah butir
Instrumen jawaban tertutup	→	Banyak
Instrumen jawaban terbuka	→	Sedikit

Dalam praktiknya, tidak ada batasan yang pasti tentang jumlah butir yang "banyak" atau "sedikit". Umumnya, jumlah butir sebanyak 35-50 dianggap cukup memadai jika kita menggunakan jenis instrumen tertutup dalam sekali pengukuran, dengan asumsi waktu tersedia 100-120 menit. Sedangkan jika menggunakan instrmen

terbuka, biasanya digunakan pertanyaan/pernyataan sebanyak 5-10 butir untuk jumlah waktu yang sama.

- Faktor reliabilitas instrumen. Berdasarkan banyak penelitian, jumlah butir instrument, khususnya instrumen dengan jawaban tertutup, jumlah butir yang digunakan dalam pengukuran berpengaruh terhadap reliabilitas atau konsistensi hasil pengukuran. Semakin banyak jumlah butir instrumen, maka reliabilitas cenderung meningkat mendekati angka 1,00 secara logaritmik. Artinya, penambahan butir cenderung meningkatkan angka reliabilitas sampai pada jumlah butir tertentu, tetapi peningkatannya cenderung melambat mulai pada jumlah butir tertentu. Umumnya, butir instrumen sebanyak 40-50 butir sudah cukup memadai untuk mencapai hasil pengukuran yang konsisten. Kaitan jumlah butir dengan angka reliabilitas ini akan dibahas pada bab berikutnya dari buku ini.

6. Menyusun Kisi-Kisi dan Butir Instrumen

Langkah terakhir dalam perencanaan penyusunan instrumen adalah menyusun kisi-kisi dan butir instrumen. Kisi-kisi merupakan format berbentuk tabel yang dijadikan sebagai batasan (*frame*) penyusunan instrumen. Kisi-kisi umumnya dinyatakan dalam format tabel dengan kolom-kolom tertentu yang minimal memuat tujuan pengukuran, aspek dan sub aspek yang diukur, butir pertanyaan/pernyataan instrumen, beserta jawabannya. Dalam format yang lebih lengkap, kisi-kisi dapat juga memuat informasi tentang pedoman pemberian skor dan karakteristik butir instrumen, seperti tingkat vaiditas dan reliabilitas instrumen, atau tingkat kesukaran dan daya pembeda butir soal tes.

Contoh kisi-kisi instrumen pengukuran aspek kognitif berbentuk tes, adalah sebagai berikut:

Tabel 4.1. Contoh kisi-kisi tes

Tujuan pengukuran	Sub aspek	Bentuk soal	Butir soal	Alternative dan kunci jawaban
Untuk mengukur kemampuan peeserta didik dalam mengurangkan bilangan pecahan berpenyebut sama	C2 C3	Tes PG	1. Jumlah dari ¾ - ¼ adalah: 2. Adi membeli kue, Kue tersebut dibagi menjadi 5 potong Setelah itu, 1 potong diberikan ke Ani dan 2 potong diberikan ke Badu. Berapa bagian kah sisa potongan kue yang ada dengan Adi ?	1* ½ 5/8 4/8 6/4 3/5 1/5 2/5* 4/5 5/5
Untuk mengukur kemampuan peserta didik dalam menyebutkan nama pendiri organisasi Muhammadiyah	C1	Tes isian singkat	1. Organisasi Muhammadiyah didirikan oleh …..	KH. Ahmad Dahlan
Untuk mengukur adanya sikap menghargai orang lain pada peserta didik	A3	Wawancara	1. Bagaimana pendapat kamu jika tempat duduk kamu bersebelahan dengan orang yang berbeda suku dengan kamu? 2. Bahasa seperti apa yang kamu gunakan pada saat kamu memanggil orangtuamu di rumah?	1. Tidak keberata 2. Mengguna kan bahasa yang halus
Untuk mengukur kemampuan	P1	Tes praktik	Tirukan suara azan (setelah	Peserta tes menirukan

peserta didik menirukan suara azan			saya contohkan)	suara azan dengan lafaz yang benar

Dari contoh pada tabel di atas, tampak bahwa sangat penting untuk menjaga kesesuaian antara tujuan pengukuran dan aspek serta sub aspek hasil belajar, dengan susunan kalimat pada butir pertanyaan/pernyataan. Salah satu fungsi tabel kisi-kisi instrumen adalah menjaga keseuaian tersebut. Dengan penyusunan kisi-kisi instrumen, evaluator lebih yakin bahwa butir-butir instrumen akan mengukur apa yang seharusnya diukur secara akurat.

Selanjutnya, format tabel kisi-kisi instrumen dapat dikembangkan sesuai dengan keadaan dan kebutuhan, misalnya menggunakan kolom yang lebih banyak dan lengkap.

7. Analisis Kualitas Instrumen

Analisis kualitas instrumen diperlukan untuk memperoleh instrumen yang handal. Analisis ini dilakukan segera setelah butir-butir soal tersusun. Sebagai bagian dari pengukuran tidak langsung, maka pengukuran hasil belajar memerlukan instrumen yang handal, sehingga hasil pengukurannya akurat dan dapat dipercaya. Beberapa uji kehandalan yang umum dilakukan adalah uji validitas dan reliabilitas instrumen. Sedangkan pada instrumen berbentuk tes, juga dilakukan uji kehandalan berupa analisis daya pembeda dan tingkat kesukaran butir soal tes. Secara lebih rinci, kehandalan instrumen akan dibahas pada bab berikutnya.

Sebelum instrumen pengukuran digunakan atau disimpan untuk keperluan mendatang, maka instrumen tersebut harus melalui pengujian kualitas instrumen. Pengujian kualitas instrumen umumnya meliputi:

a. Uji pakar atau *peer review*

Uji pakar atau *peer review* merupakan upaya untuk memperoleh instrumen yang handal secara teoretik. Secara teoretik, dengan pengujian pakar maka kualitas instrumen dapat ditingkatkan akurasinya. Pengujian oleh pakar (*judge expert*) atau oleh penilaian sejawat (*peer review*) dilakukan terhadap butir-butir instrumen, berguna sebagai pembanding atau *second opinion* terhadap kualitas instrumen yang kita susun. Asumsi dasar yang digunakan adalah, semakin banyak pakar atau teman sejawat yang ikut menelaah instrumen yang kita susun, maka semakin kecil kekeliruan yang mungkin terjadi. Sehingga, biasanya digunakan minimal sebanyak 2 (dua) orang pakar atau teman sejawat sebagai reviewer.

Pakar dan teman sejawat yang dijadikan sebagai reviewer instrumen haruslah orang yang dianggap benar-benar faham tentang evaluasi pembelajaran, atau minimal telah memiliki pengalaman yang memadai dalam pengukuran hasil belajar. Untuk itu dapat digunakan para dosen, atau guru-guru senior yang memiliki bidang yang sama dengan evaluator.

Analisis pakar atau teman sejawat, dapat dilakukan secara umum dan terbuka, tetapi dapat pula menggunakan format tertentu atau lembar validasi yang telah disediakan. Keuntungan menggunakan format analisis umum dan terbuka adalah memberi kebebasan kepada pakar atau teman sejawat untuk memberikan masukan dari berbagai sudut pandang, terutama secara kualitatif. Tentu saja koreksi umum dan terbuka ini dinyatakan secara deskriptif kualitatif, sehingga tidak dapat digunakan jika evaluator menginginkan adanya skor terhadap hasil analisis. Sedangkan keuntungan analisis instrumen menggunakan lembar validasi

dengan format tertentu adalah agar analisis, koreksi, dan masukan para pakar lebih terarah dan dapat diberikan skor dengan pendekatan kuantitatif. Kedua format tersebut pada dasarnya dapat digunakan, yang penting secara substansi analisis yang dilakukan mengacu pada kualitas komponen pembentuk instrumen tersebut. Komponen dimaksud adalah kesesuaian butir instrumen dengan tujuan pengukuran dan karakteristik peserta tes/responden, kualitas dan kesesuaian bahasa yang digunakan, kejelasan petunjuk, dan kejelasan pertanyaan atau pernyataan butir instrumen. Khusus bagi instrumen yang dimaksudkan akan menghasilkan skor hasil pengukuran, maka perlu dianalisis pula kejelasan arah jawaban serta pedoman pemberian skornya. Kadangkala evaluator menggunakan format gabungan, yakni format analisis kuantitatif dan kualitatif.

Contoh format evaluasi instrumen yang diberikan kepada pakar atau teman sejawat adalah:

LEMBAR VALIDASI INSTRUMEN PENGUKURAN

Nama instrumen : Tes Hasil Belajar mata pelajaran IPA
Penyusun/evaluator : Udin, S,Pd
Bentuk instrumen : Tes isian singkat
Jumlah butir : 25
Kisi-kisi : Terlampir
Petunjuk pengisian :

1. Berikan tanda contreng atau silang pada kolom skor yang dianggap paling sesuai dengan keadaan instrumen yang dianalisis.
2. Hitung-hitung rata-rata berdasarkan jumlah skor (tanpa pembobotan)

3. Buat kategorisasi hasil analisis berdasarkan skor rata-rata tersebut.

No	Aspek yang dianalisis					
A.	**Kesesuaian butir instrumen dengan:**	**1**	**2**	**3**	**4**	**5**
1.	Tujuan pembelajaran/standar kompetensi					
2.	Tujuan pengukuran					
3.	Aspek hasil belajar kognitif (C1, C2, C3)					
4.	Karakteristik peserta tes/responden					
B.	**Bahasa**					
1.	Menggunakan bahasa Indonesia yang baik dan benar					
2.	Menggunakan bahasa yang mudah difahami					
3.	Menggunakan bahasa yang tegas/tidak menimbulkan penafsiran ganda					
4.	Kejelasan petunjuk pengerjaan soal					
C.	**Kunci jawaban dan pemberian skor**					
1.	Ketepatan/kesesuaian kunci jawaban dengan pertanyaan					
2.	Kejelasan pedoman pemberian skor					
3.	Obyektivitas pemberian skor					
	JUMLAH SKOR					
	RATA-RATA SKOR					
	KATEGORI KUALITAS INSTRUMEN*) Baik, Cukup, Kurang (Coret yang tidak perlu)					

Komentar dan catatan revisi:

……

……

……

……

*) Catatan: Kategorisasi rata-rata skor:

Rata-rata skor	kategori
’00 - 1,67	kurang
1,68 – 3,30	cukup
3,31 – 5,00	baik

Palangka Raya, ……..

Validator/reviewer,

……..

Format di atas merupakan contoh yang dapat dikembangkan sesuai dengan kebutuha analisis instrumen.

Hasil dari pengujian oleh pakar atau teman sejawat, dapat digunakan oleh evaluator untuk menentukan butir-butir instrumen yang baik dan butir mana yang harus direvisi atau bahkan dibuang.

8. Ujicoba di Lapangan dan Analisis Butir

Ujicoba di lapangan merupakan upaya untuk memperoleh instrumen yang handal secara empirik. Kadangkala evaluator tidak hanya membutuhkan pengujian pakar atau *peer review* saja, akan tetapi juga memerlukan ujicoba lapangan untuk memperoleh bukti empirik dan menambah keyakinan tentang kualitas butir-butir instrumen. Pengujian dilakukan dengan cara memberikan butir-butir instrumen kepada sekelompok

peserta atau responden ujicoba yang memiliki karakteristik sama atau relevan dengan kelompok yang menjadi target pengukuran.

Jika pengukurannya menggunakan sampel, maka kelompok ujicoba bisa diambil dari anggota populasi yang tidak terpilih sebagai sampel pengukuran. Dalam konteks pengukuran hasil belajar, kelompok ujicoba biasanya diambil dari kelompok yang berbeda dari kelompok yang menjadi target pengukuran, misalnya pada kelas parallel pada sekolah yang sama atau berbeda. Kata kunci pada pemilihan kelompok ujicoba adalah adanya kesamaan atau relevansi karakteristik antara kelompok ujicoba dan kelompok target pengukuran.

Umumnya evaluator juga menghindari untuk melakukan ujicoba pada kelompok yang sama dengan kelompok target pengukuran, karena memungkinkan terjadinya *carry-out effect. Carry-out effect adalah* terjadinya pembiasan skor hasil pengukuran karena peserta tes atau responden sudah pernah mengerjakan dan menjawab butir-butir instrumen yang sama atau mirip, sehingga skornya cenderung meningkat.

Jumlah peserta kelompok ujicoba umumnya di atas 30 orang. Berdasarkan beberapa penelitian, jumlah peserta ujicoba di bawah 30 cenderung menghasilkan hasil ujicoba yang kurang baik dan tidak konsisten. Penambahan jumlah peserta ujicoba sampai dengan jumlah tertentu, juga akan turut meningkatkan koefisien reliabilitas instrumen.

Dari hasil ujicoba di lapangan, evaluator akan menemukan butir-butir mana dari instrumen tersebut, yang paling bagus untuk digunakan, harus direvisi, atau harus dibuang.

9. Revisi Instrumen Jika Dibutuhkan

Revisi butir-butir instrumen dibutuhkan jika berdasarkan hasil uji pakar atau *peer review,* terdapat butir-butir instrumen yang kurang baik. Jika evaluator melakukan uji pakar dan ujiocba di lapangan, maka hasil analisis keduanya dapat digunakan sebagai dasar untuk melakukan revisi. Misalkan dari hasil ujicoba instrumen di lapangan diperoleh beberapa butir yang memiliki validitas rendah, maka butir tersebut dapat direvisi. Revisi dilakukan dengan cara menelaah kembali pertanyaan (stem) dan jawaban butir instrumen tersebut, baik dari sisi bahasa, kejelasan perintah, tingkat kesukaran, dan sebagainya.

Setelah dilakukan revisi terhadap beberapa butir instrumen yang dianggap kurang baik, maka butir instrumen tersebut dapat digunakan sebagai bagian dari keseluruhan atau perangkat instrumen pengukuran.

10. Pengadministrasian Instrumen

Setelah instrumen selesai disusun, dan telah diperoleh butir-butir instrumen yang dianggap cukup baik, maka evaluator dapat melakukan penataan administrasi terhadap instrumen tersebut. Pilihannya adalah, mungkin instrumen tersebut akan segera digunakan, atau mungkin pula disimpan dan diarsipkan dalam jangka waktu tertentu untuk digunakan pada masa mendatang.

Pengadministrasian instrumen dimaksudkan sebagai langkah teknis yang dilakukan evaluator sebelum instrumen digunakan. Instrumen yang baik harus diadministrasikan pula secara baik, sehingga memudahkan evaluator dan peserta tes/responden saat digunakan.Termasuk dalam pengadministrasian instrumen antara lain:

- Pengetikan instrumen secara rapi
- Penggandaan instrumen sesuai dengan jumlah yang dibutuhkan.

- Penyiapan sistem distribusi instrumen.
- Penyiapan sistem pengumpulan kembali instrumen ke evaluator.
- Penyiapan sistem atau cara penskoran dan penilaian.
- Pengarsipan instrumen dan hasil pengukuran serta penilaian ke dalam bank soal.

BAB V
MENGUJI KEHANDALAN INSTRUMEN

A. Konsep Kehandalan Instrumen pada Teori Skor Klasik

Sebagaimana telah dipaparkan pada bagian sebelumnya, pengukuran hasil belajar termasuk ke dalam pengukuran tidak langsung. Ciri utama pengukuran tidak langsung adalah bahwa evaluator hanya mengamati dan mengukur gejala atau respon yang timbul dari obyek pengukuran. Respon tersebut kemudian diberikan skor atau dinilai sesuai dengan tujuan pengukuran. Skor atau nilai hasil pengukuran respon tersebutlah yang dianggap mewakili ukuran dari obyek yang diukur. Dengan demikian, hasil pengukuran tidak langsung masih mengandung kelemahan, dan memiliki kemungkinan pembiasan serta kekeliruan yang lebih tinggi dibandingkan dengan hasil pengukuran langsung.

Skor hasil pengukuran tidak langsung masih mengandung kekeliruan, pembiasan, dan kesalahan. Naga (1992) menyatakan bahwa skor hasil pengukuran hasil belajar masih bersifat probalistik karena mengandung unsur kekeliruan. Dengan kata lain, skor hasil hasil belajar terdiri dari skor sebenarnya (*True score*) dan skor kekeliruan (*Error*), yang dapat dilambangkan dalam persamaan berikut :

$X = T + \varepsilon$

X = skor hasil pengukuran atau pengamatan

T = *True* atau skor sebenarnya

ε = *Error*

Berdasarkan persamaan tersebut, jika seorang peserta tes atau responden memperoleh skor 80 dari hasil suatu pengykuran, maka skor 80 tersebut belum tentu menggambarkan kemampuan sebenarnya dari responden. Banyak kemungkinan kombinasi skor T dan Ɛ yang mungkin terjadi, misalnya:

80 = 70 + 10……..(1)

80 = 90 + (-10) ……..(2).

80 = 79 + 1 ……..(3).

Pada persamaan (1) di atas, jika seorang peserta tes atau responden memperoleh skor 80, maka ada kemungkinan kemampuan sebenarnya adalah 70, tetapi karena terdapat skor kekeliruan sebesar 10 maka skor yang diperoleh peserta didik atau skor hasil pengukurannya adalah 80. Kemungkinan berbeda terjadi pada persamaan (2), yang mana kemampuan sebenarnya dari peserta tes adalah 90, tetapi karena terdapat skor kekeliruan sebesar -10 maka skor yang diperoleh peserta didik atau skor hasil pengukurannya adalah 80. Ini berarti, jika seorang peserta tes memperoleh skor 80, maka akan terdapat tak terhingga kemungkinan pasangan kombinasi skor T dan Ɛ. Pada persamaan (3), skor hasil pengukuran 80, mendekati skor T yang merupakan kemampuan sebenarnya dari peserta tes, dengan skor kekeliruan hanya sebesar 1. Persamaan (3) merupakan gambaran tentang hasil pengukuran yang kita harapkan, yakni skor hasil pengukuran tidak jauh berbeda dengan skor hasil pengamatan.

Dalam pengukuran hasil belajar, tujuan utama evaluator adalah mencari skor *true* T. tetapi karena skor T tidak dapat diamati (laten, tersembunyi), maka evaluator hanya dapat mengukur dan menghasilkan skor X. Tantangan utama dalam pengukuran tidak langsung adalah, menghasilkan sebaran skor X yang paling mendekati skor T. Dengan kata lain, tantangan utama evaluator adalah meminimalkan skor kekeliruan atau *error*. Jika diusahakan skor kekeliruan atau

error mendekati nol (Ɛ≈0), maka persamaan X = T + Ɛ akan mendekati X = T + 0, sehingga skor hasil pengamatan akan hampir sama dengan angka *true* skor. Dengan kata lain, jika dapat diusahakan Ɛ≈0, maka akan terjadi skor hasil amatan mendekati *true* skor, atau X≈T. Artinya, dengan mengusahakan skor kekeliruan yang sekecil mungkin, maka skor hasil pengamatan yang kita peroleh akan mampu menggambarkan kemampuan sebenarnya dari peserta didik. Contohnya adalah pada persamaan (3) di atas. Masalahnya adalah, skor X adalah hasil pengamatan, sehingga dapat kita amati skornya. Sedangkan skor T dan Ɛ tidak dapat kita amati, tetapi secara teoretis dapat dikendalikan, antara lain dengan mengusahakan penggunaan instrumen yang handal.

Terdapat beberapa kemungkinan terjadinya kesalahan pengukuran, sehingga terjadi pembiasan skor hasil pengukuran. Kesalahan pengukuran menyebabkan skor hasil pengamatan (X) semakin menjauh dari skor *true* (T) dan skor kekeliruan (Ɛ) juga semakin besar. Artinya, kesalahan pengukuran akan menyebabkan menurunnya tingkat akurasi hasil pengukuran, sehingga skor yang kita peroleh tidak menggambarkan kemampuan sebenarnya dari peserta tes atau responden.

Kesalahan pengukuran terjadi karena adanya beberapa kekurangan, yang disebabkan oleh beberapa faktor, yakni kesalahan pada : (1) Instrumen pengukuran yang digunakan, (2), Kesalahan cara penggunaan pengukuran, (3). Kesalahan pada situasi pengukuran. Beberapa faktor penyebab tersebut dapat berdiri sendiri maupun terkombinasi satu sama lain. Kesalahan nomor (2) dan (3) relatif dapat diatasi dengan cara membuat petunjuk penggunaan instrumen dengan memperhatikan kondisi psikologis peserta tes/responden. Sedangkan kemungkinan kesalahan yang disebabkan instrumen dapat diminimalisir dengan cara menyusun instrumen yang handal. Dengan demikian, kehandalan instrumen yang digunakan evaluator, sangat menentukan kualitas dan akurasi skor hasil pengukuran.

Untuk memperoleh instrumen yang handal, evaluator harus mengusahakan tercapainya (1). Validitas instrumen secara teoretis, (2). Validitas instrumen secara empiris, (3). Reliabilitas instrumen, (4). Untuk instrumen berbentuk tes, evaluator harus mengusahakan daya pembeda yang baik dan tingkat kesukaran butir soal tes dalam taraf sedang.

Gambaran tentang jenis pengujian yang dibutuhkan untuk memperoleh instrumen yang handal digambarkan pada tabel berikut:

Tabel 5.1. Jenis pengujian kehandalan instrumen yang dibutuhkan

Jenis dan bentuk instrumen	Validitas teoretis	Validitas empiris	Reliabilitas	Daya pembeda	Tingkat kesukaran
A. Tes tertulis					
1. PG	X	X	X	X	X
2. Isian singkat	X	X	X	X	X
3. Menjodohkan	X	X	X	X	X
4. Sebab akibat	X	X	X	X	X
5. Uraian	X	X	X	-	X
B. Tes lisan	X	X	X	-	-
C. Tes praktik	X	X	X	-	-
D. Non Tes					
1. Angket tertutup	X	X	X	-	-
2. Angket terbuka	X	-	-	-	-
3. Pedoman wawancara	X	-	-	-	-
4. Pedoman dokumentasi	X	-	-	-	-
5. Pedoman observasi	X	-	-	-	-

Dari tabel di atas dapat difahami bahwa umumnya uji validitas teoretis dibutuhkan untuk semua jenis instrumen, tetapi validitas empiris dibutuhkan bagi instrumen yang hasil pengukurannya dapat dinyatakan dalam bentuk angka atau skor. Demikian pula halnya dengan pengujian reliabilitas. Sedangkan pengujian daya pembeda dan tingkat kesukaran umumnya hanya dibutuhkan untuk instrumen berbentuk tes. Perbedaan kebutuhan tersebut disebabkan oleh sifat jawaban tes yang memiliki kemungkinan jawaban benar dan salah, sehingga memungkinkan evaluator untuk membedakan

peserta tes yang pintar (yang banyak menjawab benar) dan yang kurang pintar (yang sedikit menjawab benar).

Upaya untuk memperoleh instrumen yang handal, dilakukan seara simultan. Pada saat menyusun instrumen, pertama-tama evaluator berusaha untuk mencapai validitas teoretis dengan beberapa cara. Setelah diperoleh intrumen yang memiliki butir-butir valid secara teoretis, evaluator kemudian berusaha mencapai validitas empiris dan reliabilitas, melalui ujicoba instrumen kepada kelompok peserta ujicoba. Untuk instrumen berupa tes, ujicoba tersebut juga dimaksudkan untuk mengetahui daya pembeda dan tingkat kesukaran butir tes.

Ujicoba instrumen dimaksudkan untuk mempelajari karakteristik instrumen dan butir-butir instrumen pada saat digunakan. Hasil ujicoba kemudian dianalisis menggunakan teknik tertentu, biasanya menggunakan software yang telah banyak tersedia di pasaran. Dari hasil ujicoba instrumen, evaluator dapat menentukan butir-butir insrumen yang handal sesuai dengan kebutuhan.

Upaya untuk memperoleh instrumen yang handal tersebut, akan dibahas pada bagian berikut ini:

B. Validitas

1. Makna Validitas

Validitas berasal dari *valid,* yang artinya sah atau absah. Validitas dapat diartikan sebagai tingkat keabsahan. Validitas instrumen diartikan sebagai tingkat keabsahan instrumen tersebut sebagai alat ukur. Dengan kata lain, suatu instrumen disebut memiliki validitas, jika instrumen tersebut dapat berfungsi sebagai alat ukur yang sah, atau sebagai alat ukur yang tepat. Instrumen yang valid adalah instrumen yang mampu mengukur secara tepat apa yang seharusnya diukur (Callahan & Logan, 2021). Itulah sebabnya validitas juga bisa dimaknai sebagai tingkat ketepatan instrumen sebagai

alat ukur. Artinya, semakin tepat instrumen tersebut berfungsi sebagai alat ukur, maka berarti semakin valid instrumen tersebut berfungsi sebagai alat ukur.

Contoh instrumen yang valid sebagai alat ukur adalah:

- Termometer yang digunakan sebagai alat ukur suhu.
- Meteran yang digunakan untuk mengukur panjang meja
- Altometer yang digunakan sebagai alat ukur ketinggian.
- Timbangan yang digunakan sebagai alat ukur berat badan.

Sedangkan contoh instrumen yang tidak valid sebagai alat ukur adalah:

- Meteran yang digunakan untuk mengukur berat badan.
- Jengkalan tangan yang digunakan untuk mengukur panjang meja.
- Telapak tangan yang digunakan sebagai alat ukur suhu tubuh.
- Tongkat yang digunakan sebagai alat ukur jarak.

Dari dua bagian contoh di atas, tampak bahwa validitas juga mengacu pada kecocokan antara instrumen yang digunakan sebagai alat ukur, dengan karakteristik obyek yang akan diukur. Meteran merupakan instrumen yang valid digunakan untuk mengukur panjang meja, tetapi jelas tidak valid jika digunakan untuk mengukur berat badan. Demikian pula sebaliknya, timbangan merupakan instrumen yang valid digunakan sebagai alat ukur berat badan, tetapi jelas tidak valid jika digunakan untuk mengukur panjang meja. Dengan demikian, validitas instrumen juga berarti sebagai tingkat ketepatan instrumen tersebut dalam mengukur obyek ukur yang sesuai.

Beberapa contoh di atas adalah instrumen yang umumnya digunakan dalam pengukuran langsung. Dalam pengukuran langsung, umumnya instrumen telah disepakati dan distandarisasi secara internasional. Contohnya adalah meteran. Ukuran atau skala yang digunakan sebagai patokan atau standar dalam meteran, merupakan skala yang telah disepakati secara internasional. Contoh lainnya adalah termometer, yang mana skala atau ukuran-ukuran yang digunakan telah disepakati dan distandarisasi secara internasional.

Hal yang berbeda terjadi pada pengukuran tidak langsung. Sebagaimana telah dijelaskan pada bagian awal bab ini, pengukuran hasil belajar termasuk ke dalam pengukuran tidak langsung. Instrumen dalam pengukuran tak langsung, tidak dapat distandarisasi secara luas maupun internasional. Hal ini terjadi karena sifat obyek pengukuran yang laten/tersembunyi, sehingga evaluator hanya bisa mengukur gejalanya saja. Faktor kedua adalah, hasil pengukuran tidak langsung bersifat relatif, karena hasilnya yang dapat berubah-ubah sesuai dengan keadaan yang terjadi saat pengukuran. Ada ketergantungan antara hasil pengukuran menggunakan instrumen tertentu, dengan karakteristik peserta pengukuran. Keterkaitan antara keduanya dinyatakan sebagai ciri khas teori skor klasik (Naga, 1992).

Hasil pengukuran yang masih relatif, menyebabkan keharusan bagi evaluator untuk memiliki instrumen yang benar-benar handal, antara lain harus memiliki validitas yang tinggi.

Validitas instrumen dapat dibedakan menjadi validitas teoretis dan validitas empiris. Pengelompokkan ini didasarkan pada perbedaan perlakuan pada saat penyusunan instrumen. Hal itu akan dibahas sebagai berikut:

2. Validitas Teoretis

Validitas teoretis bermakna sebagai ukuran yang menyatakan tingkat ketepatan instrumen sebagai alat ukur, yang diperoleh dari analisis teoretis. Artinya, pengujian validitas instrumen dilakukan secara teoretis, dan umumnya kualitatif. Pengujian validitas teoretis yang ditempuh oleh validitor antara lain berupa uji pakar (*judge expert*) atau review oleh teman sejawat (*peer review*).

Validitas teoretis kadangkala juga disebut sebagai validitas ini (*content validity).* Validitas isi diartikan sebagai ketepatan pertanyaan atau pernyataan butir-butir instrumen untuk mengukur apa yang seharusnya diukur. Dengan kata lain, validitas isi adalah tingkat ketepatan instrumen dilihat dari isi instrumennya.

Validitas teoretis instrumen dapat dicapai oleh validator dengan beberapa cara berikut ini:

a. Menyusun butir instrumen sesuai dengan tujuan pengukuran

   Kesesuaian antara butir instrumen dengan tujuan pengukuran, merupakan salah satu kunci utama untuk mencapai instrumen yang valid secara teoretis. Evaluator harus memiliki keyakinan secara kuat bahwa setiap butir pertanyaan atau pernyataan dalam instrumen memang ditujukan untuk mengukur hasil belajar tertentu sebagaimana tercantum dalam tujuan pengukuran. Setiap tujuan pengukuran harus terwakili dalam satu atau beberapa butir instrumen. Jika tujuan pengukurannya adalah untuk mengetahui kemampuan peserta tes dalam melakukan pembagian pecahan murni, maka butir tes yang digunakan harus menanyakan tentang kemampuan melakukan pembagian pecahan murni, tidak tercampur dengan butir-butir tes yang menanyakan

tentang kemampuan pembagian pada pecahan campuran atau kemampuan mengalikan pecahan. Jika tujuan pengukurannya adalah mengukur kemampuan peserta tes dalam menjelaskan penyebab terjadinya hujan, maka butir-butir tes yang disusun seharusnya menanyakan tentang penyebab terjadinya hujan, bukan penyebab terjadinya banjir.

Untuk meneliti relevansi pertanyaan atau pernyataan setiap butir instrumen dengan tujuan pengukurannya masing-masing, dibutuhkan ketelitian dan kehati-hatian evaluator dalam menyusun butir-butir instrumen. Pilihan kata dan kalimat yang digunakan dalam butir instrumen harus tegas, jelas, dan singkat, sehingga benar-benar relevan dengan tujuan pengukuran.

b. Menyusun butir sesuai dengan karakteristik aspek dan sub aspek hasil belajar yang akan diukur.

Selain relevansi dengan tujuan pengukuran, validitas teoretis juga dicapai dengan cara menyusun butir-butir yang sesuai dengan karakteristik peserta tes atau responden. Evaluator harus mampu memperhitungkan karakteristik peserta tes atau responden, sehingga pola respon mereka dapat diprediksi. Beberapa pertanyaan yang harus diperhatikan tentang karakteristik responden ini antara lain:

- Apakah butir-butir pertanyaan atau pernyataan dalam instrumen, cocok dengan tahapan perkembangan belajar peserta tes/responden ?. Penggunaan kata dan kalimat dalam pertanyaan instrumen, haruslah bisa difahami. Kadangkala peserta tes atau responden tidak menjawab pertanyaan dalam instrumen secara benar,

karena mereka tidak memahami pertanyaan tersebut secara benar.

Pilihan bentuk instrumen juga dipengaruhi oleh tahapan perkembangan belajar peserta tes. Misalnya, instrumen berbentuk angket terbuka tidaklah cocok diberikan kepada peserta didik kelas III SD. Demikian pula penggunaan angket tertutup untuk mengukur sikap peserta didik kelas II SD, akan menyebabkan kesulitan mereka mengisi angket karena kesulitan membaca teks.

- Apakah butir-butir pertanyaan dalam instrumen telah sesuai dengan aspek dan sub aspek hasil belajar yang akan diukur? Misalnya, jika sub aspek hasil yang akan diukur adalah kemampuan memahami (C2), maka butir-butir tes yang disusun hanya mengukur kemampuan memahami, bukan kemampuan menerapkan (C3) atau menganalisa (C4).
- Apakah butir-butir instrumen telah menggambarkan keseluruhan materi atau bahan yang akan diukur?. Penting bagi evaluator untuk memastikan bahwa butir-butir instrumen yang disusun, memiliki relevansi yang kuat dengan materi yang hendak diukur, dan untuk apa materi atau bahan tersebut diberikan kepada responden.
- Seberapa sulit pertanyaan dalam instrumen untuk dijawab ?. Untuk instrumen berbentuk tes hasil belajar, maka evaluator harus memprediksi tingkat kesukaran butir tes, sehingga tidak ada butir soal tes yang terlalu sulit untuk dijawab oleh peserta tes. Dalam pengukuran hasil belajar, sangat dianjurkan untuk menggunakan butir-butir tes dengan tingkat kesukaran sedang (Naga, 1992).

- Dalam penyusunan tes hasil belajar, perlu dikaji tentang seberapa lama jarak waktu antara penggunaan instrumen dengan pemberian tes?. Jarak waktu yang terlalu lama antara keduanya, bisa menyebabkan pembiasan hasil pengukuran. Artinya, ada kemungkinan peserta tes mampu menjawab tes dengan benar karena perkembangan tahapan belajarnya, bukan karena mereka menguasai materi pelajaran. Selain itu, jarak waktu yang terlalu lama juga menyebabkan peserta tes telah lupa tentang materi yang diajarkan.

c. Membuat kisi-kisi instrumen

   Menyusun butir-butir instrumen melalui pembuatan kisi-kisi instrumen sangat dianjurkan. Sebagaimana telah dijelaskan pada bagian sebelumnya, kisi-kisi berfungsi sebagai pemandu dalam penyusunan instrumen. Kisi-kisi instrumen akan membantu mengarahkan evaluator untuk mengukur dengan baik, sehingga validitas teoretis instrumen dapat ditingkatkan.

d. Melakukan uji pakar atau review teman sejawat.

   Uji pakar atau review oleh teman sejawat, merupakan upaya evaluator untuk memperoleh pandangan lain tentang kualitas instrumen yang disusunnya. Adanya pendapat, kritik dan saran dari para pakar dan teman sejawat, akan membantu meningkatkan ketelitian evaluator dalam penyusunan instrumen. Pada penyusunan butir-butir instrumen yang jumlahnya banyak, kadangkala evaluator berkurang tingkat ketelitiannya, sehingga memerlukan pendapat orang lain. Selain itu, tidak semua evaluator memahami dengan baik ilmu evaluasi dan teori pengukuran, sehingga membutuhkan pendapat

ahli dan uji pakar. Uji pakar atau *peer review* yang intensif akan meningkatkan validitas teoretis instrumen.

3. Validitas Empiris

Validitas empiris bermakna sebagai ukuran yang menyatakan tingkat ketepatan instrumen sebagai alat ukur, yang diperoleh dari pengujian instrumen secara empiris di lapangan. Pengujian validitas empiris dilakukan dari analisis terhadap hasil ujicoba instrumen terhadap kelompok ujicoba. Pendekatan yang digunakan lebih bersifat kuantitatif, karena skor hasil ujicoba dianalisis menggunakan rumus-rumus statistika. Validitas empiris yang diupayakan evaluator dapat berupa satu jenis atau beberapa jenis validitas berikut ini, tergantung dari kebutuhan pengukuran. Validitas tersebut dapat berupa validitas butir instrumen, maupun validitas instrumen secara keseluruhan sebagai sebuah perangkat.

Jenis validitas empiris instrumen tersebut adalah :

a. Validitas konstruksi

Validitas konstruksi merupakan jenis validitas butir instrumen, sehingga pengujiannya merupakan pengujian terhadap skor setiap butir instrumen. Validitas konstruksi adalah validitas yang ditunjukkan oleh keterkaitan antar butir instrumen sehingga terbangun konstruksi instrumen secara keseluruhan. Ibarat konstruksi sebuah gedung, bangunan gedung adalah instrumen pengukuran, dan butir-butir pertanyaan instrumen merupakan tiang-tiang dan bagian-bagian gedung yang saling menopang sehingga membentuk bangunan tersebut. Validitas konstruksi adalah kemampuan setiap butir instrumen untuk menopang satu sama lain sehingga membentuk skor hasil pengukuran yang diperoleh responden, sehingga membentuk skor total atau skor

kumulatif. Dengan kata lain, suatu butir instrumen akan memiliki validitas konstruksi yang baik, jika skor pada butir tersebut ikut berkontribusi terhadap skor total hasil pengukuran.

Agar butir-butir instrumen tersebut saling menopang satu sama lain dalam membentuk bangunan instrumen, maka butir-butir tersebut haruslah unidimensi. Butir-butir unidimensi bermakna butir-butir tersebut memiliki keterkaitan dimensi pengukuran atau indikator yang relevan satu sama lain. Jika butir-butir unidimensi, maka skor kumulatif yang diperoleh peserta tes dapat menggambarkan kemampuan peserta tes secara utuh.

Dalam praktiknya, validitas konstruksi dapat digambarkan oleh koefisien korelasi antara skor setiap butir instrumen terhadap skor kumulatif atau skor total yang diperoleh peserta tes atau responden. Dengan asumsi bahwa skor setiap butir dan skor totalnya adalah data jenis interval dan memenuhi syarat uji statistika parametrik, maka rumus yang digunakan untuk menguji validitas konstruksi adalah rumus korelasi *product moment* sebagaimana ditulis oleh Glass & Hopkins (1984) sebagai berikut:

$$r = \frac{N \sum XY - \sum X \cdot \sum Y}{\sqrt{\left\{N \sum X^2 - (\sum X)^2\right\} \left\{N \sum Y^2 - (\sum Y)^2\right\}}}$$

Keterangan:

r : Koefisien korelasi butir yang diuji terhadap skor total
X : Skor butir instrumen yang diuji validitasnya
Y : Skor total dari keseluruhan butir instrumen
N : Jumlah peserta tes atau responden

Untuk melakukan pengujian validitas konstruksi yang dimiliki setiap butir instrumen, maka dari hasil ujicoba instrumen, evaluator harus melakukan perhitungan koefisien korelasi antara skor setiap butir terhadap skor totalnya. Koefisien korelasi tersebut kemudian dibandingkan dengan tabel r pada taraf signifikansi tertentu (umumnya digunakan taraf signifikansi 5%), dan dengan derajat bebas (N-2), dengan kriteria penarikan simpulan sebagai berikut :

Butir instrumen disebut valid jika r > r tabel pada taraf signifikansi dan derajat bebas tertentu. Jika diperoleh r ≤ r tabel, maka butir instrumen tersebut dianggap tidak valid secara konstruksi.

Saat ini, terdapat banyak *software* komputer di pasaran, yang dapat digunakan oleh evaluator sebagai alat bantu perhitungan. Dengan menggunakan *software,* evaluator cukup hanya memasukkan skor-skor setiap butir hasil ujicoba instrumen, kemudian komputer secara otomatis menampilkan koefisien korelasi dan simpulan valid tidaknya butir-butir tersebut. Beberapa *software* yang populer digunakan sebagai alat bantu analisis butir instrumen antara lain ANATES, ANABUT, dan SPSS.

Sebagai contoh, berikut ini diberikan gambaran tentang pengujian validitas konstruksi butir soal tes secara manual. Tes terdiri dari 10 butir pertanyaan berbentuk Pilihan Ganda (PG), yang diujicobakan kepada 25 orang peserta ujicoba. Setiap jawaban benar diberi skor 1, dan jawaban salah diberi skor 0. Data hasil ujicoba tersebut adalah sebagai berikut:

Tabel 5.2. Contoh data hasil ujicoba tes pilihan ganda

Kode	**Skor tiap butir**										**Total**
Peserta	1	2	3	4	5	6	7	8	9	10	
1	1	1	1	0	1	1	0	1	1	0	7
2	0	1	0	1	1	1	0	1	1	1	7
3	1	1	1	1	1	1	1	1	1	0	9
4	1	1	0	0	1	1	1	1	0	0	6
5	1	0	0	0	1	1	1	1	1	0	6
6	1	1	1	1	1	1	1	1	1	0	9
7	0	1	1	1	0	0	1	1	1	1	7
8	1	1	1	0	0	0	1	1	1	0	6
9	1	1	0	1	1	1	0	1	0	0	6
10	1	1	1	1	1	0	0	1	0	1	7
11	1	0	1	1	1	1	1	0	1	0	7
12	1	1	1	1	1	0	1	1	0	0	7
13	0	0	1	1	0	1	1	0	0	1	5
14	1	1	0	0	0	1	1	1	1	0	6
15	1	1	0	0	1	1	0	1	0	0	5
16	0	1	1	1	1	1	1	0	1	0	7
17	1	1	1	1	1	0	0	1	0	0	6
18	1	1	1	1	1	1	1	1	1	0	9
19	1	0	1	1	1	1	1	0	0	1	7
20	1	1	1	1	1	0	1	1	1	0	8
21	1	1	1	0	1	1	1	1	0	0	7
22	1	1	1	1	1	0	1	0	1	0	7
23	1	1	1	1	0	1	0	1	0	1	7
24	1	0	0	1	0	1	1	0	0	0	4
25	1	1	1	1	1	1	1	1	0	1	9
Jumlah	21	20	18	18	19	18	18	19	13	7	171

Jika dari tabel di atas kita ingin menguji validitas konstruksi butir soal nomor 1, maka kita anggap sebaran skor butir soal nomor 1 adalah sebaran skor X, dan sebaran skor totalnya adalah Y. Skor X dan Y kemudian kita cari koefisien korelasinya, dengan terlebih dahulu membuat dan melengkapi tabel kerja analisis korelasi sebagai berikut:

Tabel 5.3. Contoh tabel kerja analisis validitas konstruksi butir X

Kode peserta	X	X^2	Total (Y)	Y^2	XY
1	1	1	7	49	7
2	0	0	7	49	0
3	1	1	9	81	9
4	1	1	6	36	6
5	1	1	6	36	6
6	1	1	9	81	9
7	0	0	7	49	0
8	1	1	6	36	6
9	1	1	6	36	6
10	1	1	7	49	7
11	1	1	7	49	7
12	1	1	7	49	7
13	0	0	5	25	0
14	1	1	6	36	6
15	1	1	5	25	5
16	0	0	7	49	0
17	1	1	6	36	6
18	1	1	9	81	9
19	1	1	7	49	7
20	1	1	8	64	8
21	1	1	7	49	7

22	1	1	7	49	7
23	1	1	7	49	7
24	1	1	4	16	4
25	1	1	9	81	9
Jumlah	21	21	171	1209	145

Dari tabel di atas diketahui:

$\Sigma X = 21$
$\Sigma X^2 = 21$
$\Sigma Y = 171$
$\Sigma Y^2 = 1209$
$\Sigma XY = 145$
$N = 25$

Selanjutnya koefisien korelasi antara skor butir ke-1 (X) terhadap skor totalnya adalah:

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

$$r = \frac{25\,(145) - (21)(171)}{V\,((25(21) - (21)^2)\,((25(1209) - (171)^2)}$$

$$r = \frac{3.625 - 3.591}{V\,((525) - (441))\,((30.225) - (29.241)}$$

$$r = \frac{34}{V\,(84)\,(984)}$$

$$r = \frac{34}{V\,(82.656)}$$

$$r = \frac{34}{287{,}50}$$

$r = 0{,}12$ (dibulatkan 2 desimal)

Ini berarti bahwa koefisien korelasi antara skor butir ke-1 (X) terhadap skor total (Y) adalah r = 0,12. Koefisien ini kemudian dibandingkan dengan tabel r dengan derajat bebas (N-2) atau (25-2=23) dan taraf signifikansi 5%. Tabel r lengkap dapat dilihat di lampiran.

Dari tabel r tersebut, diketahui r tabel = 0,3961 sebagaimana potongan tabel r berikut (lihat bagian yang dilingkari. Bagian ini merupakan pertemuan antara baris derajat bebas (DF) = 23 dan kolom taraf signifikansi 5% atau 0,05).

= n-2	0,1	0,05	0,02	0,0
	r 0,005	r 0,05	r 0,025	r 0,
1	0,9877	0,9969	0,9995	0,99
2	0,9000	0,9500	0,9800	0,99
3	0,8054	0,8783	0,9343	0,95
4	0,7293	0,8114	0,8822	0,91
5	0,6694	0,7545	0,8329	0,87
6	0,6215	0,7067	0,7887	0,83
7	0,5822	0,6664	0,7498	0,79
8	0,5494	0,6319	0,7155	0,76
9	0,5214	0,6021	0,6851	0,73
10	0,4973	0,5760	0,6581	0,70
11	0,4762	0,5529	0,6339	0,68
12	0,4575	0,5324	0,6120	0,66
13	0,4409	0,5140	0,5923	0,64
14	0,4259	0,4973	0,5742	0,62
15	0,4124	0,4821	0,5577	0,60
16	0,4000	0,4683	0,5425	0,58
17	0,3887	0,4555	0,5285	0,57
18	0,3783	0,4438	0,5155	0,56
19	0,3687	0,4329	0,5034	0,54
20	0,3598	0,4227	0,4921	0,53
21	0,3515	0,4132	0,4815	0,52
22	0,3438	0,4044	0,4716	0,51
23	0,3365	0,3961	0,4622	0,50
24	0,3297	0,3882	0,4534	0,49
25	0,3233	0,3809	0,4451	0,48
26	0,3172	0,3739	0,4372	0,47

Gambar 5.1. Potongan tabel r

Jika dibandingkan, maka r yang diperoleh, yakni r = 0,12, nilainya lebih kecil dari r tabel, atau berlaku r < r tabel, sehingga disimpulkan bahwa butir ke-1 bukan merupakan butir yang valid secara konstruksi.

Dengan cara dan langkah seperti di atas, kita dapat melakukan perhitungan koefisien korelasi butir ke-2, ke-3, sampai dengan butir ke-10. Berikut ini diberikan gambaran tentang hasil perhitungan koefisien korelasi setiap butir tes di atas terhadap skor totalnya, sebagai berikut:

Tabel 5.4. Tabel ringkasan koefisien korelasi butir terhadap skor total

No butir	Nilai r	r tabel	Simpulan
1	0,12	0,391	Tidak valid
2	0,42	0,391	Valid
3	0,56	0,391	Valid
4	0,35	0,391	Tidak valid
5	0,45	0,391	Valid
6	-0,01	0,391	Tidak valid
7	0,20	0,391	Tidak valid
8	0,30	0,391	Tidak valid
9	0,39	0,391	Tidak valid
10	0,08	0,391	Tidak valid

Dari hasil analisis pada contoh di atas, ternyata hanya terdapat 3 butir soal tes yang valid, yakni butir soal nomor 2,3, dan 5. Selain nomor tersebut, butor soal dianggap tidak valid. Untuk butir-butir dengan koefisien korelasi yang tidak terlalu jauh berbeda dengan kriteria t tabel, evaluator dapat menggunakan butir tersebut dengan terlebih dahulu melakukan revisi terhadap butir soal. Revisi dapat dilakukan dengan pendekatan kualitatif, dengan meneliti kembali langkah-langkah pada validitas teoretis yang telah ditempuh evaluator.

Hasil dan simpulan yang sama juga akan diperoleh jika menggunakan *software* pada komputer. Komputer secara otomatis menghitung dan menampilkan hasil pengujian.. Bahkan pada beberapa *software,* tampilan hasil analisis dilengkapi dengan simpulan.

b. Validitas kriterium

Validitas kriterium termasuk ke dalam jenis validitas perangkat instrumen, karena pengujiannya dilakukan terhadap skor total atau jumlah skor dari keseluruhan butir. Validitas kriterium adalah tingkat ketepatan instrumen jika dibandingkan dengan kriteria tertentu. Kriteria tersebut umumnya berupa skor hasil pengukuran yang telah ada dan dianggap sebagai hasil pengukuran yang standar dan akurat. Dengan demikian, validitas kriterium diperoleh dengan cara membandingkan skor hasil ujicoba terhadap skor tertentu yang dijadikan kriteria. Misalnya kita ingin mengetahui validitas tes hasil belajar matematika untuk peserta didik kelas VIII. Skor total hasil ujicoba tersebut dapat kita korelasikan dengan skor tes hasil belajar terdahulu pada peserta didik yang sama. Sebagai kriteria, kita dapat gunakan hasil tes sebelumnya, nilai raport, dan semacamnya. Syarat utama kriterium, adalah skor hasil pengukuran yang dianggap standar, akurat, dan konsisten menggambarkan hasil belajar peserta tes atau responden yang sama.

Keuntungan penggunaan validitas kriterium adalah bahwa pengujian validitasnya kita lakukan terhadap skor total, tidak harus menguji korelasi setiap butir sehingga menghemat perhitungan. Selain itu, jika skor hasil ujicoba instrumen langsung dapat digunakan sebagai skor akhir, selama nantinya diperoleh simpulan bahwa instrumen yang diujicobakan memiliki validitas kriterium yang

tinggi. Sedangkan kelemahannya, evaluator harus mencari skor kriterium yang tepat, dapat dipercaya, dan konsisten pada peserta tes atau responden yang sama. Kadangkala, perilaku dan hasil belajar peserta tes atau responden telah mengalami perubahan karena perkembangan psikologis mereka, sehingga agak sulit dijadikan sebagai kriterium.

Validitas kriterium umumnya bisa digunakan sebagai cara memperoleh instrumen yang handal, jika evaluator telah memiliki data yang akurat tentang hasil belajar peserta tes/responden yang akan diukur. Ketika evaluator akan menyusun dan mengembangkan suatu instrumen yang benar-benar baru, maka pendekatan ini agak sulit digunakan.

Untuk menguji validitas kriterium, evaluator dapat mencari koefisien korelasi antara skor hasil ujicoba instrumen (X) terhadap skor kriterium (Y), dengan menggunakan rumus korelasi *product moment* dan kriteria penarikan simpulan sebagaimana telah dipaparkan pada bagian sebelumnya.

Sebagai contoh, seorang evaluator ingin mengetahui validitas instrumen berupa angket minat terhadap mata pelajaran Biologi pada 33 orang peserta didik kelas X SMA. Pada 3 bulan sebelumnya, seorang peneliti juga telah mengukur minat terhadap mata pelajaran Biologi pada peserta didik yang sama. Angket yang digunakan peneliti tersebut tekah diujicoba sebelumnya sehingga ia memperoleh butir-butir angket yang handal, akurat, dan konsisten. Dengan demikian, evaluator dapat menggunakan skor hasil angket minat terdahulu tersebut sebagai kriterium (Y). Sebaran skor hasil pengukuran menggunakan angket oleh evaluator (X), dan sebaran skor kriterium (Y) adalah sebagai berikut:

Tabel 5.5: Contoh data pengujian validitas kriterium

Kode peserta	X	Y
1	32	33
2	35	35
3	40	44
4	37	50
5	23	20
6	25	32
7	40	38
8	35	33
9	34	45
10	45	40
11	41	40
12	42	38
13	36	34
14	37	35
15	45	42
16	36	30
17	37	35
18	38	39
19	40	36
20	42	37
21	45	40
22	42	44
23	46	43
24	45	47

25	35	32
26	33	34
27	41	40
28	35	31
29	36	32
30	37	35
31	38	36
32	35	34
33	37	38

Keterangan:

X : Skor hasil angket

Y : Skor kriterium

Jika dari tabel di atas kita ingin menguji validitas kriterium angket minat, maka kita anggap sebaran skor total pada hasil ujicoba angket minat tersebut adalah sebaran skor X, dan sebaran skor kriteriumnya adalah Y. Skor X dan Y kemudian kita cari koefisien korelasinya, dengan terlebih dahulu membuat dan melengkapi tabel kerja analisis korelasi sebagai berikut:

Tabel 5.6. Contoh tabel kerja analisis validitas kriterium

Kode peserta	X	Y	X^2	Y^2	XY
1	32	33	1024	1089	1056
2	35	35	1225	1225	1225
3	40	44	1600	1936	1760
4	37	50	1369	2500	1850
5	23	20	529	400	460
6	25	32	625	1024	800
7	40	38	1600	1444	1520

8	35	33	1225	1089	1155
9	34	45	1156	2025	1530
10	45	40	2025	1600	1800
11	41	40	1681	1600	1640
12	42	38	1764	1444	1596
13	36	34	1296	1156	1224
14	37	35	1369	1225	1295
15	45	42	2025	1764	1890
16	36	30	1296	900	1080
17	37	35	1369	1225	1295
18	38	39	1444	1521	1482
19	40	36	1600	1296	1440
20	42	37	1764	1369	1554
21	45	40	2025	1600	1800
22	42	44	1764	1936	1848
23	46	43	2116	1849	1978
24	45	47	2025	2209	2115
25	35	32	1225	1024	1120
26	33	34	1089	1156	1122
27	41	40	1681	1600	1640
28	35	31	1225	961	1085
29	36	32	1296	1024	1152
30	37	35	1369	1225	1295
31	38	36	1444	1296	1368
32	35	34	1225	1156	1190
33	37	38	1369	1444	1406
Jumlah	1245	1222	47839	46312	46771

Dari tabel di atas diketahui :

$\Sigma X = 1245$

$\Sigma X^2 = 47839$

$\Sigma Y = 1222$
$\Sigma Y^2 = 46312$
$\Sigma XY = 46771$
$N = 33$

Selanjutnya koefisien korelasi antara skor total X terhadap skor total kriteria Y adalah :

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

$$r = \frac{33\,(46.771) - (1.245)(1.222)}{V\,((33(47.836) - (1.245)^2)\,((33(46.312) - (1.222)^2)}$$

$$r = \frac{1.543.443 - 1.521.390}{V\,((1.578.588) - (1.550.025))\,((1.528.296) - (1.493.284)}$$

$$r = \frac{22.053}{V\,(28.563)\,(35.012)}$$

$$r = \frac{22.053}{V\,(1{,}000047756)}$$

$$r = \frac{22.053}{31623,5317}$$

r = 0,70 (dibulatkan 2 desimal)

Ini berarti bahwa koefisien korelasi antara skor total hasil ujicoba angket (X) terhadap skor kriterium (Y) adalah r = 0,70. Koefisien ini kemudian

dibandingkan dengan tabel r dengan derajat bebas (N-2) atau (33-2=31) dan taraf signifikansi 5%, sehingga diperoleh r tabel = 0,3440. Dengan demikian, r > r tabel, sehingga disimpulkan bahwa skor hasil ujicoba angket minat tersebut berkorelasi tinggi dengan skor kriterium. Artinya, angket minat yang diujicobakan meniliki validitas kriterium yang tinggi. Dengan kata lain, skor hasil ujioba angket minat tersebut telah mampu menggambarkan minat peserta didik kelas X SMA secara konsisten.

Kelebihan pendekatan validitas kriterium ini adalah, evaluator dapat menggunakan skor hasil ujicoba instrumen sebagai skor akhir hasil pengukuran, jika korelasi antara skor hasil ujicoba terhadap skor kriterium cukup tinggi. Dengan kata lain, evaluator tidak perlu lagi melakukan pengukuran untuk mengetahui skor hasil belajar yang akan diukur, karena skor hasil ujioba instrumen sudah dapat menggambarkan kemampuan yang akan diukur.

c. Validitas prediktif

Validitas prediktif termasuk ke dalam validitas perangkat instrumen, karena pengujiannya dilakukan terhadap skor total atau jumlah skor dari keseluruhan butir.

Validitas prediktif adalah tingkat ketepatan atau kemampuan instrumen untuk memprediksi hasil belajar pada masa mendatang. Pada saat digunakan dan diujicobakan, validitas prediktif suatu instrumen belum dapat ditentukan. Validitas ini baru dapat dihitung beberapa tahun kemudian, dengan cara menganalisis relevansi dan ketepatan hasil pengukurannya dengan kriterium tertentu. Pada dasarnya, jenis validitas prediktif mirip dengan validitas kriterium. Perbedaannya adalah pada

kriteriumnya, karena yang digunakan sebagai kriterium bukan skor hasil pengukuran pada saat itu, tetapi adalah skor hasil pengukuran pada masa mendatang. Jika skor hasil pengukuran suatu instrumen berkorelasi positif secara signifikans dengan skor hasil pengukuran pada masa datang, maka berarti instrumen tersebut memiliki validitas prediktif yang tinggi.

Validitas prediktif biasanya digunakan untuk mencari insttumen yang mampu memprediksi hasil belajar seseorang pada masa mendatang. Jika ditemukan instrumen yang memiliki validitas prediktif yang tinggi, maka berarti instrumen tersebut dapat digunakan untuk memprediksi hasil belaar pada orang atau kelompok lain dengan karakteristik yang sama. Misalnya kita ingin menguji validitas prediktif tes ujian masuk perguruan tinggi yang digunakan beberapa tahun yang lalu, maka kita dapat menggunakan indeks prestasi kumulatif mahasiswa saat ini kriterium. Jika terdapat korelasi positif yang signifikans antara kedua skor tersebut, maka berarti tes ujian masuk perguruan tinggi yang digunakan beberapa tahun yang lalu tersebut memiliki validitas prediktif.

Teknik analisis yang digunakan untuk mencari validitas prediktif suatu instrumen, adalah dengan mencari korelasi antara skor hasil pengukuran instrumen tersebut (X) pada masa lalu, terhadap skor kriterium (Y) yang digunakan saat ini. Cara analisis dan kriteria penarikan simpulannya sama dengan analisis korelasi untuk mencari validitas kriterium.

C. Reliabilitas

Relliabilitas adalah kemampuan suatu instrumen untuk menghasilkan skor hasil pengukuran yang konsisten, ajeg, dan relatif tetap. Itulah sebabnya reliabilitas disebut juga sebagai tingkat ketetapan atau tingkat keajegan. Instrumen yang baik dan reliabel adalah instrumen yang menghasilkan hasil pengukuran yang tetap meskipun digunakan pada waktu, tempat, dan keadaan berbeda. Pengukuran yang reliabel adalah pengukuran yang menghasilkan hasil yang relatif sama ketika diulang kepada kelompok yang sama atau sejenis (Callahan & Logan, 2021).

Ukuran reliabilitas dapat dipandang dari dua sisi, yakni indeks reliabilitas dan koefisien reliabilitas. Indeks reliabilitas hanya dapat didekati secara konseptual, karena adanya berbagai keterbatasan, sehingga angkanya tidak dapat dihitung. Sedangkan koefisien reliabilitas dapat didekati secara praktis menggunakan analisis statistika.

Jika dikaitkan dengan konsep pengukuran tidak langsung dengan formula skor hasil pengukuran $X = T + Ɛ$ sebagaimana telah dipaparkan sebelumnya, maka reliabilitas dimaknai sebagai perbandingan antara variansi skor hasil pengukuran X terhadap variansi skor *true* (T). (Naga, 1997, Sax, 1980, Bulkani, 2020). Nilai perbandingan inilah yang dinamakan indeks reliabilitas. Dengan demikian, indeks reliabilitas dinyatakan sebagai $\rho = \sigma^2_X/\sigma^2_T$ di mana ρ adalah indeks reliabilitas, σ^2_X adalah variansi skor hasil pengukuran, dan σ^2_T adalah variansi skor *true* yang menggambarkan kemampuan sebenarnya dari peserta tes atau responden. Jika variansi skor σ^2_X semakin mendekati variansi skor σ^2_T, maka semakin besar nilai reliabilitas ρ.

Indeks reliabilitas ρ berada pada kisaran 0,00-1,00. Indeks ρ akan mencapai nilai maksimum pada $\rho = 1{,}00$, yakni jika $\sigma^2_X = \sigma^2_T$ atau variansi skor hasil pengukuran X sama dengan variansi skor *true* T. Pada keadaan $\sigma^2_X = \sigma^2_T$, maka kesalahan pengukuran atau *error* pada persamaan $X = T + Ɛ$ adalah sama dengan nol, yang berarti sebagai kondisi ideal yang

diharapkan dalam pengukuran. Sedangkan indeks ρ sama atau mendekati nol terjadi jika variasi skor hasil pengukuran σ^2_X mendekati nol. Naga (1997) berpendapat bahwa instrumen yang reliabel adalah instrumen yang menghasilkan skor hasil pengukuran yang akurat dengan skor kekeliruan atau *error* sekecil-kecilnya.

Jika indeks reliabilitas kita artikan sebagai $\rho = \sigma^2_X/\sigma^2_T$, maka untuk mencarinya kita harus mengetahui sebaran skor X dan sebaran skor T, sehingga variansi kedua jenis sebaran tersebut juga dapat dicari. Karena X adalah skor hasil pengukuran atau skor hasil pengamatan, tentu saja sebaran skor X kita ketahui dari hasil pengukuran menggunakan instrumen tertentu. Akan tetapi, sebaran skor *true* atau T adalah skor laten yang akan kita cari dari persamaan $X = T + \varepsilon$, sehingga skor T tidak kita dapatkan. Dengan kata lain, indeks reliabilitas ρ sebenarnya secara praktis tidak dapat kita hitung karena ketidakmampuan kita menentukan sebaran skor T. Ironinya, jika kita sudah mengetahui sebaran skor *true* T, maka kita tentunya kita tidak perlu lagi membuat instrumen yang handal untuk mencari sebaran skor X, karena skor T adalah tujuan yang kita cari dalam melakukan pengukuran.

Berbeda dengan indeks reliabilitas, maka koefisien reliabilitas melambangkan konsistensi atau keajegan hasil pengukuran dari waktu ke waktu, dari satu butir instrumen ke butir yang lain, dari satu keadaan ke keadaan lainnya, pada peserta tes atau responden yang sama. lain, dan seterusnya. Reliabilitas dalam hal ini dipandang sebagai konsistensi hasil pengukuran, terlepas dari berbagai konteks yang terkait dengan instrumen dan situasi pengukuran. Dalam pandangan umum, koefisien reliabilitas inilah yang sering disebut sebagai reliabilitas instrumen.

Sesuai dengan konsep di atas, maka reliabilitas (yang dimaksud di sini adalah koefisien reliabilitas), dapat dihitung dengan menggunakan pendekatan korelasi. Pandangan ini didasari pada asumsi bahwa dalam pengukuran berlaku :

1. Kemampuan sebenarnya dari peserta tes atau responden tentang materi yang diukur, atau kita sebut *true* atau T dalam persamaan X = T + Ɛ, merupakan skor yang tetap. Artinya, dengan menggunakan instrumen apapun yang sejenis dan tepat, maka skor T seseorang tidak akan berubah.
2. Secara statistika, sebaran skor *error* atau Ɛ pada persamaan X = T + Ɛ, akan mengikuti sebaran distribusi Normal dengan rata-rata adalah nol. Dengan demikian, semakin banyak instrumen diberikan kepada seorang peserta tes atau responden, maka skor keleliruan Ɛ akan semakin mendekati nol. Pada persamaan X = T + Ɛ, jika skor Ɛ mendekati nol, maka skor hasil pengukuran atau skor hasil pengamatn X akan semakin mendekati skor T. Implikasi dari asumsi ini adalah, skor hasil pengukuran X akan lebih mampu menggambarkan kemampuan sebenarnya dari peserta tes atau responden, jika dilakukan beberapa kali pengukuran pada peserta tes atau responden yang sama.

Asumsi 1 dan 2 di atas merupakan asumsi yang mendasari pemahaman, bahwa reliabilitas instrumen dapat dihitung dengan pendekatan korelasi antara hasil suatu pengukuran dengan hasil pengukuran lainnya yang relevan, pada peserta tes atau responden yang sama.

Sebagaimana koefisien korelasi, maka koefisien reliabilitas angkanya berada pada kisaran dari -1,00 hingga 1,00. Akan tetapi dalam konteks butir-butir instrumen yang unidimensi dan saling mendukung satu sama lain, maka koefisien reliabilitas yang diakui adalah koefisien yang positif. Semakin mendekati angka 1,00, semakin baik reliabilitas instrumen tersebut. Sebagai patokan, koefisien reliabilitas danggap baik dan cukup memadai jika diperoleh angka 0,70 ke atas, meskipun pada beberapa kasus angka di atas 0,60 masih bisa ditolerir. Naga (1997) menyarankan penggunaan koefisien reliabilitas di atas 0,75 sebagai patokan untuk cabang ilmu yang metode pengukurannya sudah baik, dan di atas 0,50 bagi

cabang ilmu yang belum begitu baik pengukurannya. Batasan tersebut juga tergantung dari pendekatan atau rumus yang digunakan dalam menghitung koefisien reliabilitas. Penggunaan rumus Spearman-Brown yang cenderung *overestimate* misalnya, akan membutuhkan batasan minimal koefisien reliabilitas yang lebih tinggi.

Sebagai patokan umum tingkat reliabilitas, dapat pula digunakan kriteria yang umumnya digunakan untuk menginterpretasi tingkat korelasi, yakni sebagai berikut:

Tabel 5.7. Tabel kriteria reliabilitas instrumen

Koefisien reliabilitas (r)	Kriteria
< 0,200	Sangat jelek
0,200 – 0,399	Jelek
0,400 – 0,599	Cukup
0,600 – 0,799	Baik
0,800 – 1,000	Sangat baik

Sekali lagi, kriteria di atas hanya dapat dijadikan sebagai patokan umum saja. Sebagaimana telah dijelaskan sebelumnya, ditolak atau diterimanya koefisien reliabilitas instrumen, juga tergantung dari cabang ilmu yang diukur, pendekatan atau rumus yang digunakan dalam perhitungan beserta asumsi-asumsi dasar yang digunakan dalam perhitungan.

Sebagian besar perhitungan koefisien reliabilitas menggunakan asumsi-asumsi dasar ilmu statistika, dengan skor hasil pengukuran berdistrbusi Normal yang akan tercapai pada jumlah pengulangan pengulangan yang sangat banyak. Pada kenyataannya, sulit bagi evaluator untuk menggunakan instrumen yang sama sebanyak-banyaknya atau tak terhingga kali kepada peserta tes atau responden yang sama. Selain karena terbatasnya waktu dan sumberdaya

lainnya, maka pemberian instrumen yang sama berulang kali kepada peserta tes juga akan memberikan efek psikologis kurang baik bagi peserta tes atau responden, sehingga justeru akan menyebabkan pembiasan hasil pengukuran.

Untuk mencari koefisien relabilitas, maka kelemahan tersebut dapat diatasi oleh evaluator antara lain dengan cara:

1. Melakukan pengulangan pengukuran menggunakan instrumen yang sama kepada peserta tes atau responden yang sama. Misalkan melakukan pengukuran sebanyak dua kali menggunakan instrumen yang sama, kemudian mengkorelasikan kedua skor hasil pengukuran tersebut. Jika ditemukan koefisien korelasi positif yang signifikans, maka itu berarti bahwa instrumen tersebut memiliki koefisien reliabilitas yang tinggi. Sebagian ahli menyebut reliabilitas jenis ini sebagai reliabilitas eksternal, karena ada unsur pembanding atau kriterium yang digunakan. Secara konseptual pendekatan ini mirip konsep validitas kriterium. Hanya yang berperan sebagai kriterium adalah hasil pengukuran berikutnya dengan menggunakan instrumen yang sama. Pendekatan ini dianggap lebih sederhana karena hanya menggunakan teknik korelasi *product moment* dari Pearson.

   Contoh analisis untuk mencari koefisien reliabilitas menggunakan pendekatan antar pengukuran 1 dan 2 menggunakan instrumen dan peserta tes atau responden yang sama adalah sebagai berikut:

   Seorang evaluator ingin mengetahui reiabilitas instrumen berbentuk tes uraian sebanyak 8 butir, yang diujicobakan kepada 30 orang peserta tes sebanyak 2 kali pengukuran atau 2 kali ujicoba. Peskoran hasil tes menggunakan rentang skor 0-5 untuk setiap butir tes. Skor hasil ujicoba pertana (X) dan skor hasil ujicoba kedua (Y) adalah sebagai berikut:

Tabel 5.8. Contoh skor hasil ujicoba tes uraian sebanyak 2 kali.

Kode Peserta	Hasil Pengukuran ke-1 (X)	Hasil Pengukuran ke-2 (Y)
1	23	20
2	32	30
3	21	20
4	19	25
5	35	30
6	32	28
7	35	38
8	30	35
9	28	27
10	26	29
11	27	29
12	25	28
13	27	30
14	36	38
15	32	33
16	28	30
17	22	35
18	10	27
19	17	15
20	20	19
21	27	28
22	31	32
23	30	34

24	23	24
25	30	30
26	28	29
27	15	14
28	18	23
29	20	25
30	32	34

Selanjutnya dapat dilakukan perhitungan koefisien korelasi antara skor X dengan skor Y dengan tabel dan langkah analisis sebagai berikut:

Tabel 5.9. Tabel kerja analisis korelasi skor hasil pengukuran 1 dan 2

Kode Peserta	X	Y	X^2	Y^2	XY
1	23	20	529	400	460
2	32	30	1024	900	960
3	21	20	441	400	420
4	19	25	361	625	475
5	35	30	1225	900	1050
6	32	28	1024	784	896
7	35	38	1225	1444	1330
8	30	35	900	1225	1050
9	28	27	784	729	756
10	26	29	676	841	754
11	27	29	729	841	783
12	25	28	625	784	700
13	27	30	729	900	810

14	36	38	1296	1444	1368
15	32	33	1024	1089	1056
16	28	30	784	900	840
17	22	35	484	1225	770
18	10	27	100	729	270
19	17	15	289	225	255
20	20	19	400	361	380
21	27	28	729	784	756
22	31	32	961	1024	992
23	30	34	900	1156	1020
24	23	24	529	576	552
25	30	30	900	900	900
26	28	29	784	841	812
27	15	14	225	196	210
28	18	23	324	529	414
29	20	25	400	625	500
30	32	34	1024	1156	1088
Jumlah	779	839	21425	24533	22627

Dari tabel kerja analisis korelasi di atas, kita dapat menghitung koefisien korelasi antara skor hasil pengukuran ke-1 (X) dengan skor hasil pengukuran ke-2 (Y) sebagai berikut:

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

$$r = \frac{30\,(22.627) - (779)(839)}{V\,((30(21.425) - (779)^2)\,((30(24.533) - (839)^2)}$$

$$r = \frac{678.810 - 653.581}{V\,((642.750) - (606.841))\,((735.990) - (703.921))}$$

$$r = \frac{25.229}{V\,(35.909)\,(32.069)}$$

$$r = \frac{25.229}{V\,(115.166.721)}$$

$$r = \frac{25.229}{33.934{,}7274}$$

r = 0,74 (dibulatkan 2 desimal)

Dari hasil perhitungan diperoleh r = 0,74 yang merupakan koefisien reliabilitas tes uraian yang diujicobakan tersebut.

2. Memberikan seperangkat tes sebanyak satu kali, yang di dalamnya mengandung beberapa kelompok butir yang saling terkait dan relevan, kemudian mencari koefisien korelasi antara satu bagian dengan bagian lainnya. Dengan demikian, peserta tes atau responden dianggap mendapatkan dua kali pengukuran menggunakan instrumen yang sama atau relevan. Teknik ini yang

digunakan untuk mencari koefisien reliabilitas menggunakan rumus Spearman-Brown. Teknik Spearman Brown menggunakan pendekatan belah-dua, yakni dengan cara membagi perangkat instrumen menjadi dua bagian yang diasumsikan mengukur hal yang sama, misalkan bagian ganjil dan bagian genap. Hasil pengukuran oleh dua bagian ini kemudian dikorelasikan untuk melihat konsistensi hasil pengukuran. Rumus Spearman-Brown yang digunakan untuk mencari koefisien reliabilitas instrumen adalah:

$$r = \frac{2\, r_{12}}{1 + r_{12}}$$ …….. (Mardapi, 2012).

r = Koefisien reliabilitas instrumen

r_{12} = Koefisien korelasi antara skor belahan 1 dengan belahan 2

Rumus ini digunakan jika evaluator yakin atau terdapat bukti bahwa butir-butir instrumen pada belahan 1 dan belahan 2 merupakan 2 bagian yang unidimensi, atau antara bagian 1 dan 2 dianggap parallel atau sejajar dan homogen. Salah satu bukti unidimensi dan kesejajaran antara bagian 1 dan 2 dapat dilihat dari kesamaan variansi skor dari kedua belahan tersebut. Untuk itu, jika diperlukan dapat dilakukan uji homogenitas variansi antara bagian 1 dan bagian 2 dari belahan instrumen.

Dasar dari pendekatan Spearman-Brown adalah korelasi antara bagian 1 dan bagian 2 dari perangkat instrumen. Jika korelasi antara bagian 1 dan bagian 2 cukup tinggi, maka koefisien reliabilitas Spearman-Brown juga tinggi. Korelasi yang tinggi terjadi jika bagian 1 dan bagian 2 dari instrumen tersebut unidimensi atau

mengukur hal yang sama. Itulah sebabnya, koefisien reliabilitas Spearman-Brown dapat digunakan untuk memprediksi atau menggambarkan konsistensi internal suatu instrumen.

Untuk menghitung koefisien reliabilitas perangkat instrumen, evaluator cukup membagi skor hasil pengukuran dari suatu instrumen tersebut menjadi 2 bagian, misalkan dibagi menjadi bagian butir ganjil dan bagian butir genap. Skor total kedua bagian ini masing-masing dijumlahkan, kemudian dicari koefisien korelasi antara keduanya sehingga diperoleh r_{12}. Koefisien korelasi antara 2 bagian tersebut kemudian dimasukkan ke dalam rumus di atas untuk menemukan koefisien reliabilitas instrumen secara keseluruhan.

Dari rumus di atas, tampak bahwa koefisien reliabilitas instrumen yang dihasilkan menggunakan pendekatan Spearman-Brown cenderung lebih tinggi, terutama karena perbandingan antara angka 2 r_{12} terhadap angka $(1 + r_{12})$ yang mana angka 2 $_{r12}$ selalu berada pada kisaran 0,00-1,00, sehingga akan selalu menghasilkan angka hasil perbandingan yang lebih tinggi. Dengan demikian, koefisien reliabilitas Spearman Brown cenderung *overestimate*, terutama jika dibandingkan dengan koefisien korelasi antar bagian instrumen. Hal itu dapat dilihat pada tabel perbandingan berikut ini:

Tabel 5.10. Perbandingan korelasi antar bagian dengan koefisien Spearman-Brown

No	Korelasi antar bagian instrumen (r_{12})	Nilai 2 r_{12}	Nilai $(1+r_{12})$	Selisih $(1+r_{12})$ dengan 2 r_{12}	Reliabilitas Spearman Brown
1	0,56	1,12	1,56	0,44	0,71
2	0,60	1,20	1,60	0,40	0,75

3	0,65	1,30	1,65	0,35	0,79
4	0,71	1,42	1,71	0,29	0,83
5	0,75	1,50	1,75	0,25	0,86
6	0,80	1,60	1,80	0,20	0,89

Dari tabel di atas tampak bahwa koefisien reliabilitas Spearman-Brown cenderung lebih tinggi dibandingkan dengan koefisien korelasi antar belahan instrumen. Hal ini disebabkan karena selisih antara angka $(1+r_{12})$ sebagai faktor pembagi atau pembilang terhadap angka 2 r_{12} sebagai penyebut, akan semakin kecil. Fenomena ini merupakan salah satu kelemahan sekaligus kritik bagi penggunaan rumus Spearman Brown. Hasil perhitungan koefisien reliabilitas menggunakan rumus Spearman-Brown harus dimaknai secara lebih hati-hati, sehingga dianggap koefisien reliabilitas sebenarnya berada di bawah hasil yang diperoleh.

Contoh perhitungan menggunakan rumus Spearman-Brown adalah sebagai berikut:

Misalnya kita akan mencari koefisien reliabilitas angket berskala Likert sebanyak 10 butir dengan skala penskoran antara 1-5, yang diujicobakan kepada 25 orang responden. Sebaran skor hasil ujicoba angket adalah sebagai berikut:

Tabel 5.11. Contoh hasil ujicoba 10 butir angket berskala Likert

Responden	1	2	3	4	5	6	7	8	9	10	Total ganjil (X)	Total genap (Y)
1	3	4	5	3	3	2	4	3	4	5	19	17
2	4	5	3	2	3	4	3	3	4	4	17	18
3	5	4	5	4	3	4	4	5	2	3	19	20
4	2	3	2	1	2	3	2	3	2	1	10	11
5	3	2	3	3	2	3	4	3	2	3	14	14

6	4	3	4	5	5	5	4	5	4	5	21	23
7	3	4	5	4	3	3	3	4	3	2	17	17
8	4	3	3	4	3	2	3	4	3	4	16	17
9	4	5	5	5	4	5	5	4	4	5	22	24
10	4	3	3	2	3	4	3	2	3	4	16	15
11	3	4	3	4	3	4	5	4	4	5	18	21
12	2	1	2	1	2	3	2	1	2	3	10	9
13	3	3	4	2	3	4	3	2	3	2	16	13
14	1	2	3	2	1	2	3	2	1	2	9	10
15	4	5	4	5	4	4	4	5	5	4	21	23
16	4	5	5	5	5	4	5	5	3	4	22	23
17	5	4	5	4	5	4	5	4	5	5	25	21
18	3	2	3	4	5	4	5	4	3	4	19	18
19	2	3	2	3	2	1	1	1	2	2	9	10
20	3	4	3	4	5	3	4	5	4	5	19	21
21	3	4	4	5	4	4	5	3	4	3	20	19
22	4	3	4	3	4	5	4	5	5	4	21	20
23	3	4	5	4	5	4	3	4	5	4	21	20
24	5	4	4	5	4	5	4	5	5	4	22	23
25	3	4	4	5	4	4	5	4	4	5	20	22

Untuk menghitung koefisien reliabilitas Spearman-Brown, maka kita terlebih dahulu mencari koefisien korelasi antara skor total bagian ganjil (X) dan bagian genap (Y) dengan tabel kerja sebagai berikut:

Tabel 5.12. Tabel kerja analisis korelasi bagian ganjil dan genap

Responden	**X**	**Y**	**X^2**	**Y^2**	**XY**
1	19	17	361	289	323
2	17	18	289	324	306
3	19	20	361	400	380
4	10	11	100	121	110

5	14	14	196	196	196
6	21	23	441	529	483
7	17	17	289	289	289
8	16	17	256	289	272
9	22	24	484	576	528
10	16	15	256	225	240
11	18	21	324	441	378
12	10	9	100	81	90
13	16	13	256	169	208
14	9	10	81	100	90
15	21	23	441	529	483
16	22	23	484	529	506
17	25	21	625	441	525
18	19	18	361	324	342
19	9	10	81	100	90
20	19	21	361	441	399
21	20	19	400	361	380
22	21	20	441	400	420
23	21	20	441	400	420
24	22	23	484	529	506
25	20	22	400	484	440
	443	449	8313	8567	8404

Untuk menghitung koefisien korelasi antara bagian ganjil dengan bagian genap, digunakan rumus dan langkah analisis korelasi *product-moment* sebagaimana telah dicontohkan pada bagian sebelumnya, yakni sebagai berikut:

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

$$r = \frac{25\,(8.404) - (443)(449)}{V\,((25(8.313-(443)^2)\,((25(8.567) - (449)^2)}$$

$$r = \frac{210.100 - 198.907}{V\,((207.825) - (196.249))\,((214.175) - (201.601)}$$

$$r = \frac{11.193}{V\,(11.576)\,(12.574)}$$

$$r = \frac{11.193}{V\,(145.556.624)}$$

$$r = \frac{11.193}{12.064,68499}$$

r = 0,93 (dibulatkan 2 desimal)

r = 0,93 merupakan koefisien korelasi antara bagian ganjil dan bagian genap instrumen yang diujicobakan. Sedangkan koefisien reliabilitas instrumen secara keseluruhan menggunakan rumus Spearman-Brown adalah sebagai berikut:

$$r = \frac{2\, r_{12}}{1 + r_{12}}$$

$$r = \frac{2\,(0{,}93)}{1 + (0{,}93)}$$

$$r = \frac{1{,}86}{1{,}93}$$

$$r = 0{,}96$$

Dengan demikian, diperoleh koefisien reliabilitas angket yang diujicobakan tersebut adalah 0,96. Koefisien reliabilitas ini cenderung lebih tinggi dari koefisien korelasi antar bagian genap dan ganjil sebagaimana telah dihitung.

3. Memberikan seperangkat tes yang didalamnya mengandung butir-butir instrumen unidimensi, sehingga instrumen tersebut tidak lagi dibelah menjadi dua bagian, tetapi dibelah menjadi n bagian yang mana n adalah jumlah butir tes. Dalam pendekatan ini, setiap butir instrumen dianggap merupakan bagian dari keseluruhan perangkat instrumen. Dengan demikian, diasumsikan bahwa setiap peserta tes atau responden mendapatkan pengukuran sebanyak n kali menggunakan instrumen yang sama atau relevan. Pendekatan ini digunakan dalam rumus Alpha-Cronbach dan rumus Kuder Richardson (KR).

   Rumus Alpha-Cronbach, digunakan jika skor hasil pengukuran tidak dapat dibelah menjadi 2 bagian, karena tidak memiliki variansi yang sama, atau tidak cukup bukti bahwa belahan skor hasil pengukuran adalah paralel dan homogen (Mardapi, 2012, Cronbach, 1985).

Rumus alpha-Cronbach yang digunakan cukup beragam bentuk, tetapi pada dasarnya sama, yakni dengan terlebih dahulu menghitung variansi skor setiap butir instrumen dan variansi skor total hasil pengukuran, kemudian membanding antara keduanya. Rumus yang digunakan antara lain adalah:

$$r_{\alpha} = \frac{k\,((s^2y - (s^21 + s^22 + \ldots + s^2k))}{(k\text{-}1)\,s^2y} \quad \ldots\ldots\text{(Mardapi, 2012, Naga, 1992)}$$

Keterangan:

r_{α} = koefisien reliabilitas alpha-Cronbach
s^2y = variansi skor total hasil pengukuran
s^2k = variansi skor butir ke-1, 2, hingga ke-k
k = banyak butir instrumen

Contoh:

Seorang evaluator ingin mengetahui koefisien reliabilitas instrumen berbentuk tes uraian sebanyak 6 butir yang diujicobakan kepada 30 orang peserta tes, dengan peskoran setiap butir tes adalah 0-5. Data hasil ujicoba sebagaimana pada tabel 5.12 berikut:

Tabel 5.13. Contoh hasil ujicoba tes uraian untuk analisis alpha-Cronbach

Kode peserta	X1	X2	X3	X4	X5	X6	Total (Y)
1	3	4	5	3	2	3	20
2	4	3	2	3	4	2	18
3	4	5	3	2	3	4	21
4	3	2	3	4	2	1	15
5	4	3	4	5	3	2	21
6	2	3	2	3	4	5	19
7	3	2	3	4	3	4	19

8	3	4	5	4	3	3	22
9	3	2	3	1	2	3	14
10	3	2	3	2	1	3	14
11	3	2	3	4	3	5	20
12	3	4	3	4	3	5	22
13	3	2	3	2	1	3	14
14	5	4	4	3	4	4	24
15	3	2	3	3	2	4	17
16	4	3	4	3	5	4	23
17	1	2	3	2	3	4	15
18	3	2	3	4	3	3	18
19	4	3	4	3	2	3	19
20	4	3	4	3	4	3	21
21	3	2	3	2	3	2	15
22	3	4	3	2	3	4	19
23	3	2	3	2	2	3	15
24	4	3	4	5	4	4	24
25	3	2	3	2	3	3	16
26	4	3	4	5	3	4	23
27	4	5	5	4	3	4	25
28	4	3	2	3	3	4	19
29	3	2	3	2	3	3	16
30	3	4	5	4	3	5	24
Jumlah							

Untuk melakukan pengujian koefisien reliabilitas menggunakan rumus alpha Cronbach, maka diperlukan tabel kerja sebagai berikut:

Tabel 5.14. Tabel kerja pengujian koefisien alpha-Cronbach

Kode peserta	X1	$X1^2$	X2	$X2^2$	X3	$X3^2$	X4	$X4^2$	X5	$X5^2$	X6	$X6^2$	Total (Y)	Y^2
1	3	9	4	16	5	25	3	9	2	4	3	9	20	400
2	4	16	3	9	2	4	3	9	4	16	2	4	18	324
3	4	16	5	25	3	9	2	4	3	9	4	16	21	441
4	3	9	2	4	3	9	4	16	2	4	1	1	15	225
5	4	16	3	9	4	16	5	25	3	9	2	4	21	441
6	2	4	3	9	2	4	3	9	4	16	5	25	19	361
7	3	9	2	4	3	9	4	16	3	9	4	16	19	361
8	3	9	4	16	5	25	4	16	3	9	3	9	22	484
9	3	9	2	4	3	9	1	1	2	4	3	9	14	196
10	3	9	2	4	3	9	2	4	1	1	3	9	14	196
11	3	9	2	4	3	9	4	16	3	9	5	25	20	400
12	3	9	4	16	3	9	4	16	3	9	5	25	22	484
13	3	9	2	4	3	9	2	4	1	1	3	9	14	196
14	5	25	4	16	4	16	3	9	4	16	4	16	24	576
15	3	9	2	4	3	9	3	9	2	4	4	16	17	289
16	4	16	3	9	4	16	3	9	5	25	4	16	23	529
17	1	1	2	4	3	9	2	4	3	9	4	16	15	225
18	3	9	2	4	3	9	4	16	3	9	3	9	18	324
19	4	16	3	9	4	16	3	9	2	4	3	9	19	361
20	4	16	3	9	4	16	3	9	4	16	3	9	21	441
21	3	9	2	4	3	9	2	4	3	9	2	4	15	225
22	3	9	4	16	3	9	2	4	3	9	4	16	19	361
23	3	9	2	4	3	9	2	4	2	4	3	9	15	225
24	4	16	3	9	4	16	5	25	4	16	4	16	24	576
25	3	9	2	4	3	9	2	4	3	9	3	9	16	256
26	4	16	3	9	4	16	5	25	3	9	4	16	23	529
27	4	16	5	25	5	25	4	16	3	9	4	16	25	625
28	4	16	3	9	2	4	3	9	3	9	4	16	19	361
29	3	9	2	4	3	9	2	4	3	9	3	9	16	256
30	3	9	4	16	5	25	4	16	3	9	5	25	24	576
Jumlah	99	343	87	279	102	368	93	321	87	275	104	388	572	11244
Kode peserta	X1	$X1^2$	X2	$X2^2$	X3	$X3^2$	X4	$X4^2$	X5	$X5^2$	X6	$X6^2$	Total (Y)	Y^2

Dari tabel kerja di atas, kita dapat menghtung variansi skor total (s^2y) dan variansi masing-masing butir tes s^21, s^22, dstnya hingga s^26, sebagai berikut:

Variansi total:

$$s^2y = \frac{11.244 - (572)^2/30}{30}$$

$$s^2y = \frac{11.244 - (327.184/30}{30}$$

$$s^2y = \frac{11.244 - 10.906,1333}{30}$$

$$s^2y = \frac{377,8667}{30}$$

$$s^2y = 11,26$$

Variansi butir:

$$s^21 = \frac{343 - (99)^2/30}{30} = 0,54$$

$$s^22 = \frac{279 - (87)^2/30}{30} = 0,89$$

$$s^23 = \frac{368 - (102)^2/30}{30} = 0,71$$

$$s^24 = \frac{321 - (93)^2/30}{30} = 1,09$$

$$s^2 5 = \frac{275 - (87)^2/30}{30} = 0{,}76$$

$$s^2 6 = \frac{388 - (104)^2/30}{30} = 0{,}92$$

Dengan demikian, koefisien reliabilitas alpha Cronbach (r_α) adalah:

$$r_\alpha = \frac{k\,((s^2 y - (s^2 1 + s^2 2 + \ldots + s^2 k))}{(k-1)\, s^2 y}$$

$$r_\alpha = \frac{6\,((11{,}26 - (0{,}54+0{,}89+0{,}71+1{,}09+0{,}76+0{,}92))}{(6-1)\,(11{,}26)}$$

$$r_\alpha = \frac{6\,((11{,}26 - (4{,}91))}{5(11{,}26)}$$

$$r_\alpha = \frac{6\,(6{,}35)}{56{,}30}$$

$$r_\alpha = \frac{38{,}10}{56{,}30}$$

$$r_\alpha = 0{,}68$$

Dengan demikian, koefisien reliabilitas tes uraian yang diujicobakan tersebut adalah r_{α} = 0,68. Koefisien reliabilitas alpha-Cronbach relatif di bawah koefisien reliabilitas sesungguhnya atau *underestimate* (Mardapi, 2012). Sebagai perbandingan, jika skor hasil tes uraian tersebut di atas kita analisis dengan teknik pembelahan ganjil-genap, kemudian dihitung koefisien reliabilitasnya menggunakan rumus Spearman-Brown, maka akan kita peroleh koefisien reliabilitas r = 0,74. Tampak bahwa koefisien reliabilitas tes uraian menggunakan rumus Speraman-Brown cenderung lebih tinggi dibandingkan dengan menggunakan rumus alpha-Cronbach.

Selanjutnya, untuk skor hasil pengukuran berbentuk dikotomik (misalnya skor 1 untuk jawaba benar dan skor 0 untuk jawaban salah pada skor tes pilihan ganda), maka rumus alpha-Cronbach dikembangkan oleh Kuder-Richardson-20 (KR-20). Rumus KR-2- kadangkala juga disebut sebagai koefisien reliabilitas α-20. Rumus KR-20 adalah sebagaimana ditulis Mardapi (2012) dan Naga (1992) sebagai berikut:

$$KR\text{-}20 = \left(\frac{k}{k-1}\right)\left(\frac{s^2y - (p_1(1-p_1) + p_2(1-p_2) + \ldots + p_k(1-p_k)}{s^2y}\right)$$

Keterangan:

KR-20	= Koefisien reliabilitas KR-20
k	= Banyak butir instrumen
s^2y	= Variansi skor total
p_k	= Proporsi jawaban benar, atau jumlah jawaban benar pada butir ke-k

Selanjutnya Kuder-Richardson juga menyusun rumus berdasarkan rata-rata proporsi jawaban benar dari para peserta tes, yang selanjutnya disebut rumus KR-21.

Perbedaan antara KR-20 dan KR-21, terletak pada pendekatan perhitungan yang digunakan. KR-20 didasarkan pada proporsi atau jumlah jawaban benar dari peserta tes, sedangkan pada KR-21 digunakan rata-rata jumlah jawaban benar.

Rumus KR-21 adalah:

$$KR\text{-}21 = \left(\frac{k}{k-1}\right)\left(\frac{s^2y - k\,(p_x)(1-p_x)}{s^2y}\right)$$ .....(Mardapi, 2012, Naga, 1992)

Keterangan:

KR-21 = Koefisien reliabilitas KR-21
k = Banyak butir instrumen
s^2y = Variansi skor total
p_x = Rata-rata proporsi subyek yang menjawab benar, atau jumlah seluruh nilai proporsi dibagi jumlah butir dibagi jumlah peserta tes

Contoh:

Seorang evaluator akan mencari koefisien reliabilitas KR-20, pada instrumen tes pilihan ganda dengan skor dikotomik, sebanyak 10 butir soal yang diujicobakan pada 25 orang. Data hasil ujiocba adalah sebagai berikut :

Tabel 5.15. Contoh data hasil ujicoba tes PG

Kode Peserta	skor Tiap butir										Total
	1	2	3	4	5	6	7	8	9	10	
1	1	1	1	0	1	1	0	1	1	0	7
2	0	1	0	1	1	1	0	1	1	1	7
3	1	1	1	1	1	1	1	1	1	0	9
4	1	1	0	0	1	1	1	1	0	0	6
5	1	0	0	0	1	1	1	1	1	0	6
6	1	1	1	1	1	1	1	1	1	0	9
7	0	1	1	1	0	0	1	1	1	1	7
8	1	1	1	0	0	0	1	1	1	0	6
9	1	1	0	1	1	1	0	1	0	0	6
10	1	1	1	1	1	0	0	1	0	1	7
11	1	0	1	1	1	1	1	0	1	0	7
12	1	1	1	1	1	0	1	1	0	0	7
13	0	0	1	1	0	1	1	0	0	1	5
14	1	1	0	0	0	1	1	1	1	0	6
15	1	1	0	0	1	1	0	1	0	0	5
16	0	1	1	1	1	1	1	0	1	0	7
17	1	1	1	1	1	0	0	1	0	0	6
18	1	1	1	1	1	1	1	1	1	0	9
19	1	0	1	1	1	1	1	0	0	1	7
20	1	1	1	1	1	0	1	1	1	0	8
21	1	1	1	0	1	1	1	1	0	0	7
22	1	1	1	1	1	0	1	0	1	0	7
23	1	1	1	1	0	1	0	1	0	1	7
24	1	0	0	1	0	1	1	0	0	0	4
25	1	1	1	1	1	1	1	1	0	1	9
Jumlah	21	20	18	18	19	18	18	19	13	7	171

Dari tabel di atas, dapat disusun tabel kerja analisis sebagai berikut:

Tabel 5.16. Tabel kerja pengujian koefisien reliabilitas dengan KR-20

Kode Peserta	skor tiap Butir										Total y	y^2
	1	2	3	4	5	6	7	8	9	10		
1	1	1	1	0	1	1	0	1	1	0	7	49
2	0	1	0	0	1	1	1	0	1	1	6	36
3	1	1	1	1	1	1	1	1	1	0	9	81
4	0	1	1	0	1	1	1	0	0	0	5	25
5	1	1	0	0	1	1	1	1	1	0	7	49
6	1	1	1	1	1	1	1	0	1	0	8	64
7	0	1	1	0	1	0	1	1	1	1	7	49
8	0	1	1	0	0	0	1	1	1	0	5	25
9	1	1	1	1	1	1	1	1	0	0	8	64
10	0	1	1	0	1	0	0	1	0	1	5	25
11	1	1	1	1	1	1	1	0	1	0	8	64
12	1	1	1	1	1	0	1	1	0	0	7	49
13	0	1	1	0	1	1	1	0	0	1	6	36
14	0	1	0	0	1	1	1	1	1	0	6	36
15	1	1	0	0	1	1	0	1	0	0	5	25
16	0	1	1	1	1	1	1	0	1	0	7	49
17	0	1	1	0	1	0	0	1	0	0	4	16
18	1	1	1	1	1	1	1	1	1	0	9	81
19	0	0	1	0	1	1	1	0	0	1	5	25
20	1	1	1	0	1	0	1	1	1	0	7	49
21	0	1	1	0	1	1	1	0	0	0	5	25
22	0	1	1	1	1	0	1	0	1	0	6	36
23	0	1	1	1	0	1	0	0	0	0	4	16
24	1	1	1	0	1	1	1	0	0	0	6	36
25	0	1	1	1	1	1	1	0	0	1	7	49
Jumlah	11	24	21	10	23	18	20	13	13	6	159	1059
Jawaban benar	11	24	21	10	23	18	20	13	13	6		159
p	0.44	0.96	0.84	0.4	0.92	0.72	0.8	0.52	0.52	0.24		6,36
(1-p)	0.56	0.04	0.16	0.6	0.08	0.28	0.2	0.48	0.48	0.76		
p(1-p)	0.25	0.04	0.13	0.24	0.07	0.2	0.16	0.25	0.25	0.182		1.776

Dari tabel di atas, dapat dihitung beberapa nilai yang dibutuhkan untuk rumus KR-20, yakni:

Variansi skor total:

$$s^2y = \frac{1.059 - (159)^2/25}{25} = 1,910$$

Jumlah $p_k(1-p_k)$ = 0,25+0,04+ …+0,182 = 1,776

Maka koefisien reliabilitas KR-20 adalah:

$$KR\text{-}20 = \left(\frac{10}{10-1}\right)\left(\frac{1,910 - (1,776)}{1,910}\right)$$

$$KR\text{-}20 = \left(\frac{10}{9}\right)\left(\frac{0,134}{1,910}\right)$$

KR-20 = (1,11111) (0,07015707)

KR-20 = 0,08 (dibulatkan 2 desimal)

Dengan demikian, diperoleh koefisien reliabilitas KR-20 = 0,08, yang termasuk kategori reliabilitas sangat jelek.

Sedangkan jika menggunakan rumus KR-21, akan diperoleh koefisien reliabilitas sebagai berikut:

p_x = Rata-rata proporsi subyek yang menjawab benar, atau jumlah seluruh nilai proporsi dibagi jumlah butir.

p_x = 6,36/10
p_x = 0,636
$(1- p_x)$ = 0,364

Jika dimasukkan ke rumus KR-21, maka diperoleh:

$$KR\text{-}21 = \left(\frac{k}{k-1}\right)\left(\frac{s^2y - k\,(p_x)(1-p_x)}{s^2y}\right)$$

$$KR\text{-}21 = \left(\frac{10}{10-1}\right)\left(\frac{1{,}91 - (10)\,(0{,}636)(0{,}364)}{1{,}91}\right)$$

$$KR\text{-}21 = \left(\frac{10}{10-1}\right)\left(\frac{1{,}91 - (2{,}315)}{1{,}91}\right)$$

$$KR\text{-}21 = \left(\frac{10}{10-1}\right)\left(\frac{-0{,}405}{1{,}91}\right)$$

$$KR\text{-}21 = \left(\frac{10}{9}\right)\left(-0{,}213\right)$$

$$KR\text{-}21 = \left(1{,}111\right)\left(-0{,}213\right)$$

KR-21 = -0,24 (dibulatkan 2 desimal)

Tampak di sini bahwa koefisien reliabilitas yang diperoleh adalah -0,24, yang menandakan bahwa instrumen tersebut sangat tidak reliabel. Selain itu, koefisien reliabilitas yang diperoleh dengan menggunakan KR-21 cenderung lebih rendah jika dibandingkan dengan menggunakan KR-20.

Dengan adanya beberapa rumus koefisien reliabilitas, maka evaluator diberikan beberapa pilihan untuk menggunakan salah satunya, sesuai dengan kebutuhan

pengukuran dan pertimbangan beberapa faktor. Berikut ini diberikan beberapa penjelasan tentang karakteristik masing-masing rumus koefisien reliabilitas sebagai pertimbangan evaluator sebelum menggunakannya.

Tabel 5.17. Perbandingan jenis koefisien reliabilitas berdasarkan karakteristiknya

Jenis reliabilitas	Karakteristik
Pengukuran berulang	- Digunakan jika evaluator melakukan pengukuran dua kali atau lebih menggunakan instrumen yang sama terhadap peserta yang sama. - Dapat digunakan untuk mencari koefisien reliabilitas pada hasil ujicoba tes maupun non tes,
Spearman-Brown	- Digunakan jika skor hasil pengukuran dibelah menjadi 2 bagian, misalnya belahan butir nomor ganjil dan butir nomor genap - Membutuhkan asumsi bahwa kedua belahan harus homogen, unidimensi, memiliki variansi yang tidak jauh berbeda. - Dapat menggambarkan konsistensi internal instrumen. - Dapat digunakan untuk mencari koefisien reliabilitas pada hasil ujicoba instrumen berupa tes maupun non tes. - Hasil perhitungan cenderung *overestimate*
Alpha-Cronbach	- Digunakan jika skor hasil

	pengukuran tidak dapat dibelah menjadi 2 bagian, karena tidak memiliki variansi yang sama, atau tidak cukup bukti belahan skor hasil pengukuran adalah paralel dan homogen. - Dapat digunakan untuk mencari koefisien reliabilitas pada hasil ujicoba instrumen berupa tes maupun non tes. - Dapat menggambarkan konsistensi internal instrumen. - Hasil perhitungan cenderung *underestimate* - Dapat digunakan untuk analisis reliabilitas instrumen berbentuk tes maupun non tes, khususnya pada penskoran non-dikotomik.
KR-20	- Digunakan pada instrumen dengan penskoran berbentuk dikotomik. - Perhitungan didasarkan pada proporsi atau jumlah jawaban benar pada butir instrumen.
KR-21	- Digunakan pada instrumen dengan penskoran berbentuk dikotomik. - Perhitungan didasarkan pada rata-rata proporsi atau rata-rata jawaban benar pada butir instrumen.

Terdapat beberapa faktor yang mempengaruhi tinggi rendahnya koefisien reliabilitas, antara lain jumlah butir instrumen dan jumlah peserta tes atau responden

ujicoba. Umumnya, semakin banyak jumlah peerta ujicoba, maka semakin tinggi koefisien reliabilitas. Demikian pula halnya dengan jumlah butir, semakin banyak butir instrumen, semakin tinggi koefisien reliabilitas, meskipun penambahan koefisien reliabilitas karena penambahan butir tersebut tidak secara linear (Naga, 1997, Mardapi, 2012, Bulkani, 2020).

Peningkatan koefisien reliabilitas sebagai akibat penambahan jumlah butir, dapat dihitung dengan rumus Spearman Brown sebagai berikut:

$$r' = \frac{R.r}{1 + (R-1)r} \quad \text{……(Mardapi, 2012)}$$

Dimana:

$R = (k_t + k_a)/k_a$

Keterangan :
r' = Koefisien reliabilitas setelah penambahan butir
r = Koefisien reliabilitas sebelum penambahan butir.
R = Rasio jumlah butir setelah dan sebelum penambahan butir
k_a = Banyak butir awal atau sebelum ditambahkan
k_t = Banyaknya butir yang ditambahkan.

Sebagai contoh, kita akan menambahkan sebanyak 10 butir tes terhadap perangkat tes yang semula berjumlah 30 butir, sehingga menjadi 40 butir tes. Jika koefisien reliabilitas pada awalnya adalah 0,75, maka koefisien reliabiitas setelah penambahan butir tes tersebut adalah:

$$r' = \frac{R.r}{1 + (R-1)r} \quad \text{..... (Mardapi, 2012)}$$

$$r' = \frac{(40/30).(0,75)}{1 + (40/30-1)\,(0,75)}$$

$$r' = \frac{(1,33)(0,75)}{1 + (1,33-1)\,(0,75)}$$

$$r' = \frac{0,9975}{1 + (0,33)\,(0,75)}$$

$$r' = \frac{0,9975}{1 + (0,2475)}$$

$$r' = \frac{0,9975}{1,2475}$$

r' = 0,80 (dibulatkan 2 desimal).

Dari perhitungan tampak bahwa ada peningkatan koefisien reliabilitas dari semula 0,75 menjadi 0,80 karena penambahan butir sebanyak 10 butir.

Rumus di atas juga dapat digunakan untuk memprediksi jumlah butir yang dapat kita tambahkan untuk mencapai koefisien reliabilitas yang kita inginkan. Rumus yang digunakan adalah :

$$R = \frac{r' (1-r)}{r (1-r')}$$

Keterangan:

R = Rasio jumlah butir setelah dan sebelum penambahan butir

r' = Koefisien reliabilitas setelah penambahan butir

r = Koefisien reliabilitas sebelum penambahan butir.

Misalkan kita memiliki perangkat instrumen sebanyak 30 butir dengan koefisien reliabilitas 0,75. Jika kita menginginkan peningkatan koefisien reliabilitas instrumen tersebut menjadi 0,80, maka rasio antara jumlah butir setelah dan sebelum penambahan adalah:

$$R = \frac{r' (1-r)}{r (1-r')}$$

$$R = \frac{(0,80)(1-0,75)}{(0,75)\ (1-0,80)}$$

$$R = \frac{(0,80)(0,25)}{(0,75)\ (0,20)}$$

$$R = \frac{0,20}{0,15}$$

$$R = 1,33 \text{ atau } 4/3$$

Maka jumlah butir baru yang ditambahkan (k_t) adalah :

$k_t = k_a(R-1)$

$k_t = 30\ (4/3-1)$

$k_t = 30(1/3)$

$k_t = 30(0,33)$

$k_t = 10$ butir

D. Daya Pembeda Butir Tes

Daya pembeda adalah kemampuan instrumen untuk membedakan antara responden yang menjawab benar dengan responden yang menjawab salah. Karena hal ini berkaitan dengan kemampuan menjawab benar dan salah, maka daya pembeda (DP) hanya berlaku pada instrumen berbentuk tes, khususnya untuk penskoran berbentuk dikotomik yang membutuhkan penskoran dengan nilai kebenaran mutlak. Contohnya adalah soal tes berbentuk pilihan ganda dan isian singkat. Sedangkan untuk tes dengan peskoran non-dikotomik seperti tes uraian, nilai kebenaran jawaban tes lebih bersifat relatif dan tidak mutlak sehingga perhitungan DP menjadi kurang akurat.

Dalam konteks pengukuran hasil belajar, maka daya pembeda diartikan sebagai kemampuan butir soal tes untuk membedakan antara peserta didik yang pintar dan peserta didik yang kurang pintar. Butir soal tes yang baik, haruslah merupakan butir soal yang mampu dijawab benar oleh peserta didik yang pintar, dan dijawab salah oleh peserta didik yang kurang pintar. Sebaliknya, jika butir soal tes tersebut cenderung dijawab benar oleh peserta didik yang kurang pintar dan cenderung dijawab salah oleh peserta didik

yang pintar, maka butir soal tersebut memiliki daya pembeda yang jelek.

Untuk menentukan daya pembeda atau DP suatu butir tes, maka evaluator harus membagi peserta tes menjadi 2 kelompok, yakni kelompok peserta tes yang pintar atau kelompok atas (sering disebut sebagai kelompok *upper*), dan kelompok peserta tes yang kurang pintar atau kelompok bawah (sering disebut sebagai kelompok *lower*). Pembagian kelompok ini bisa didasarkan pada :

1. Skor hasil tes dari hasil ujicoba tes. Dari skor hasil tes tersebut, evaluator dapat mengelompokkan peserta tes menjadi 2 kelompok. Kelompok atas (*upper)* adalah diambil dari peserta tes yang memperoleh skor paling tinggi. Sedangkan kelompok bawah (*lower)* diambil dari peserta tes yang memperoleh skor terendah. Umumnya, sebagai kelompok *upper* diambil sebanyak 27% peserta tes yang mendapat skor tertinggi. Sedangkan sebagai kelompok *lower* diambil sebanyak 27% peserta tes yang memperoleh skor tes terendah. Artinya, jika ada 100 orang peserta tes, maka skor hasil tesnya diurutkan dari urutan 1 sampai urutan 100. Sebagai kelompok *lower* adalah peserta tes yang berada pada urutan 1 sampai 27, sedangkan sebagai kelompok *upper* adalah peserta tes yang berada pada urutan 73 sampai 100.

   Pembagian 27% kelompok *upper* dan 27% *lower* didasarkan pada kenyataan bahwa pembagian dengan cara ini merupakan persentase yang ideal, sehingga sebagian besar evaluator menggunakan persentase tersebut. Kadangkala evaluator harus memilih persentase kelompok *upper* dan *lower* kurang atau lebih dari 27%. Masalahnya adalah, besarnya persentase kelompok *upper* dan *lower* ini berpengaruh terhadap tingkat kepercayaan kita terhadap hasil pengukuran (Naga, 1997). Persentase kelompok *upper* dan *lower* yang terlalu kecil, akan memperbesar kontras atau perbedaan antara kedua kelompok, yang juga berarti akan meningkatkan daya

pembeda butir tes. Akan tetapi,persentase kelompok *upper* dan *lower* yang terlalu kecil akan menurunkan tingkat kepercayaan kita pada hasil perhitungan karena sedikitnya sampel pesertanya. Fenomena sebaliknya terjadi jika persentase kelompok *upper* dan *lower* terlalu besar. Persentase kelompok *upper* dan *lower* yang terlalu besar akan mempersempit kontras antar kedua kelompok tersebut, yang berarti menurunkan daya pembeda butir soal dan sebaliknya akan meningkatkan kepercayaan kita karena sampel yang lebih besar. Naga (1997) menyarankan besaran persentase kelompok *upper* dan *lower* berada dalam rentangan 20% sampai 33%.

2. Data hasil tes sebelumnya. Evaluator dapat menentukan kelompok *upper* dan kelompok *lower* berdasarkan informasi atau data yang telah dimilikinya. Dalam evaluasi hsail belajar di sekolah, guru umumnya sudah mengetahui siapa peserta didik yang pintar dan siapa peserta didik yang kurang pintar. Informasi ini dapat digunakan oleh guru maupun evaluator sebagai dasar untuk menentukan kelompok *upper* dan *lower.*

Terdapat beberapa rumus yang dapat digunakan untuk menentukan daya pembeda (DP), antara lain yang paling populer adalah rumus yang mencari kontras antara kelompok *upper* dan *lower,* yakni sebagai berikut:

$$DP_i = \frac{(U_i - L_i)}{(U + L)} \quad \text{……..(Naga, 1992)}$$

Keterangan :
DP_i = Daya pembeda butir i
U_i = Jumlah kelompok *upper* yang menjawab benar butir i.
L_i = Jumlah kelompok *lower* yang menjawab benar butir i.
(U+L) = Jumlah kelompok *upper* dan *lower.*

Untuk penarikan simpulan, evaluator dapat berpatokan kepada kriteria yang dikemukakan Naga (1997), sebagai berikut:

DP < 0,20	Butir dibuang
0,20 ≤ DP < 0,30	Butir mengalami banyak revisi
0,30 ≤ DP < 0,40	Butir mengalami sedikit revisi
DP ≥ 0,40	Butir memuaskan, tanpa perlu revisi

Sebagai patokan, umumnya evaluator beranggapan bahwa daya pembeda suatu butir dianggap baik dan dapat diterima jika DP ≥ 0,30, meskipun mungkin membutuhkan beberapa revisi.

Contoh:

Seorang evaluator akan menghitung daya pembeda butir-butir tes pilihan ganda sebanyak 10 butir tes, yang diujicobakan kepada 30 orang peserta ujicoba, dengan hasil ujicoba sebagai berikut:

Tabel 5.18. Contoh data hasil ujicoba untuk menguji daya pembeda

Kode peserta	1	2	3	4	5	6	7	8	9	10	Total
1	1	1	1	1	1	0	1	1	1	1	9
2	0	0	0	0	1	0	0	0	1	1	3
3	1	0	0	1	0	1	0	0	1	0	4
4	1	1	0	1	1	1	1	1	1	0	8
5	0	1	1	0	0	1	0	0	0	1	4
6	1	0	1	0	1	0	1	0	0	0	4
7	1	0	1	0	1	0	1	1	1	0	6
8	1	1	0	0	0	1	1	0	1	0	5
9	1	1	1	1	0	1	1	1	1	0	8
10	0	1	1	1	0	0	0	1	1	1	6
11	1	1	1	1	0	0	1	0	0	0	5
12	1	0	0	1	1	1	0	0	1	1	6

13	1	1	1	1	1	1	1	0	1	0	8
14	1	0	1	1	0	0	1	1	0	0	5
15	1	1	0	1	1	1	0	0	1	0	6
16	0	1	1	0	0	0	1	0	1	1	5
17	1	1	1	1	1	0	1	1	0	0	7
18	0	1	1	1	0	0	0	1	1	1	6
19	1	0	0	1	1	1	1	0	0	0	5
20	1	1	1	0	0	0	1	1	1	0	6
21	0	0	0	0	1	0	0	0	1	1	3
22	1	1	1	0	1	0	0	1	1	1	7
23	0	0	1	1	1	0	0	1	1	0	5
24	1	0	0	0	1	0	1	0	0	1	4
25	1	1	1	1	0	0	0	1	1	0	6
26	1	0	0	1	1	0	0	1	1	1	6
27	0	1	0	1	0	0	0	0	0	1	3
28	0	1	1	0	1	1	1	1	0	1	7
29	1	1	1	1	0	1	1	1	1	0	8
30	0	1	0	0	0	0	0	0	0	1	2

Dari tabel di atas, dapat dilakukan pemeringkatan skor hasil tes dari skor terendah hingga skor tertinggi. Kemudian dilakukan pemilahan antara kelompok *upper* dan kelompok *lower*, dengan mengambil sebanyak 27% x 30 orang atau sekitar 8 orang masing-masing kelompok. Tabel di atas dapat disederhanakan menjadi tabel yang hanya memuat hasil tes kelompok *upper* dan *lower* sebagai berikut :

Tabel 5.19. Tabel kerja analisis daya pembeda butir tes

Kode awal*)	Kode baru**)	1	2	3	4	5	6	7	8	9	10	Total
	Upper											
1	1	1	1	1	1	1	0	1	1	1	1	9
4	2	1	1	0	1	1	1	1	1	1	0	8
9	3	1	1	1	1	0	1	1	1	1	0	8
13	4	1	1	1	1	1	1	1	0	1	0	8
17	5	1	1	1	1	1	0	1	1	0	0	7
22	6	1	1	1	0	1	0	0	1	1	1	7
28	7	0	1	1	0	1	1	1	1	0	1	7
29	8	1	1	1	1	0	1	1	1	1	0	8
Jumlah		7	8	7	6	6	5	7	7	6	3	
Kode awal*)	Kode baru**)	1	2	3	4	5	6	7	8	9	10	Total
	Lower											
2	1	0	0	0	0	1	0	0	0	1	1	3
3	2	1	0	0	1	0	1	0	0	1	0	4
5	3	0	1	1	0	0	1	0	0	0	1	4
6	4	1	0	1	0	1	0	1	0	0	0	4
21	5	0	0	0	0	1	0	0	0	1	1	3
24	6	1	0	0	0	1	0	1	0	0	1	4
27	7	0	1	0	1	0	0	0	0	0	1	3
30	8	0	1	0	0	0	0	0	0	0	1	2
Jumlah		3	3	2	2	4	2	2	0	3	6	

*). Kode awal adalah kode peserta tes sebelum dilakukan pemeringkatan.

**). Kode baru adalah kode peserta tes setelah pemeringkatan dan ditetapkan sebagai kelompok *upper* dan *lower*

Dari tabel di atas dapat dihitung daya pembeda masing-masing butir tes atau DP_i sebagai berikut :

$$DP_1 = \frac{7-3}{8+8} = 0{,}25$$ (perlu banyak revisi)

$$DP_2 = \frac{8-3}{8+8} = 0{,}3125$$ (perlu sedikit revisi)

$$DP_3 = \frac{7-2}{8+8} = 0{,}3125 \text{ (perlu sedikit revisi)}$$

$$DP_4 = \frac{6-4}{8+8} = 0{,}125 \text{ (butir tes dibuang)}$$

$$DP_5 = \frac{7-3}{8+8} = 0{,}25 \text{ (perlu banyak revisi)}$$

$$DP_6 = \frac{5-2}{8+8} = 0{,}1875 \text{ (butir tes dibuang)}$$

$$DP_7 = \frac{7-2}{8+8} = 0{,}3125 \text{ (perlu sedikit revisi)}$$

$$DP_8 = \frac{7-0}{8+8} = 0{,}43 \text{ (butir tes sudah memadai)}$$

$$DP_9 = \frac{6-3}{8+8} = 0{,}1875 \text{ (butir tes dibuang)}$$

$$DP_{10} = \frac{3-6}{8+8} = -0{,}1875 \text{ (butir tes dibuang)}$$

Dari hasil pengujian di atas, dapat disimpulkan bahwa butir-butir tes yang dapat digunakan adalah nomor butir tes 8. Butir tes yang dapat digunakan setelah direvisi adalah butir

nomor 1, 2,3 5, dan 7. Sedamgkan butir lainnya harus digugurkan karena memiliki daya pembeda jelek atau sangat jelek.

E. Tingkat Kesukaran Butir Tes

Tingkat kesukaran (TK) butir tes adalah angka yang menyatakan sulit atau mudahnya butir tes. Dari batasan tersebut, jelas bahwa tingkat kesukaran hanya berlaku pada instrumen berbentuk tes. Pada instrumen berbentuk non tes, tidak dikenal tingkat kesukaran, karena semua jawaban pada non tes adalah benar atau tidak ada yang salah. Sama dengan daya pembeda butir tes, maka tingkat kesukaran butir tes umumnya juga hanya berlaku pada butir tes yang jawabannya berbentuk skor dikotomik, karena pola jawaban dikotomik memiliki skor yang kebenarannya mutlak.

Tingkat kesukaran butir tes terdiri atas 3 kategori, yakni butir soal sukar, sedang, dan mudah. Butir soal disebut sukar, jika hanya sebagian kecil peserta tes yang mampu menjawab benar butir soal tersebut. Sebaliknya, butir soal tes tersebut dianggap mudah jika sebagian besar peserta tes mampu menjawab butir tes tersebut secara benar.

Untuk menghitung tingkat kesukaran butir, digunakan pendekatan proporsi sederhana, yakni banyaknya peserta tes yang menjawab benar suatu butir dibagi dengan jumlah peserta tes. Jika dikaitkan dengan pembagian kelompok *upper* dan *lower* ini, maka butir tes yang baik adalah butir tes yang dianggap mudah oleh kelompok *upper* dan dianggap sukar oleh kelompok *lower,* dan karakteristik itu hanya dimiliki oleh butir-butir tes yang termasuk dalam tingkat kesukaran sedang.

Proporsi tingkat kesukaran sukar : sedang : mudah dalam perangkat tes cukup bervariasi, tergantung dari tujuan pengukuran yang ditetapkan evaluator. Sebagian evaluator menetapkan proporsi tingkat kesukaran butir tes 20% : 50% : 30%. Ada pula yang menggunakan proporsi 20% butir tes sukar, 60% butir tes sedang, dan 20% butir tes mudah.

Sebagian ahli pengukuran menyarankan agar butir-butir tes yang digunakan dalam pengukuran hasil belajar sebaiknya hanya butir-butir dengan tingkat kesukaran sedang (Naga, 1997, Mardapi, 2012). Alasan utama dari pendapat ini adalah adanya asumsi bahwa, butir-butir tes yang sukar akan cenderung tidak bisa dijawab oleh peserta tes, dan sebaliknya butir-butir tes yang mudah cenderung dapat dijawab oleh semua peserta tes. Fenomena tersebut menghilangkan fungsi butir tes untuk membedakan peserta tes yang pintar dan kurang pintar.

Rumus yang digunakan untuk menghitung tingkat kesukaran (TK), antara lain dengan pendekatan proporsi sederhana sebagai berikut:

$$TK_i = \frac{B_i}{N} \quad \text{.....(Mardapi, 2012, Naga, 1996)}$$

Keterangan :
TK = Tingkat kesukaran butir i
B_i = Jumlah peserta tes yang menjawab benar butir i.
N = Jumlah peserta tes.

Untuk penarikan simpulan, evaluator dapat berpatokan kepada kriteria yang dikemukakan Naga (1997), yang menetapkan bahwa butir tes yang baik adalah butir tes yang memiliki tingkat kesukaran $0{,}33 \leq TK_i \leq 0{,}67$. Tingkat kesukaran butir tes termasuk sukar jika $TK_i < 0{,}33$, dan dianggap mudah jika $TK_i > 0{,}67$.

Sebagai contoh, kita akan menentukan TK terhadap data hasil ujicoba tes sebagaimana pada tabel Tabel 5.18 pada halaman sebelumnya.

Tabel 5.20. Contoh data hasil ujicoba untuk menguji tingkat kesukaran

Kode peserta	1	2	3	4	5	6	7	8	9	10	Total
1	1	1	1	1	1	0	1	1	1	1	9
2	0	0	0	0	1	0	0	0	1	1	3
3	1	0	0	1	0	1	0	0	1	0	4
4	1	1	0	1	1	1	1	1	1	0	8
5	0	1	1	0	0	1	0	0	0	1	4
6	1	0	1	0	1	0	1	0	0	0	4
7	1	0	1	0	1	0	1	1	1	0	6
8	1	1	0	0	0	1	1	0	1	0	5
9	1	1	1	1	0	1	1	1	1	0	8
10	0	1	1	1	0	0	0	1	1	1	6
11	1	1	1	1	0	0	1	0	0	0	5
12	1	0	0	1	1	1	0	0	1	1	6
13	1	1	1	1	1	1	1	0	1	0	8
14	1	0	1	1	0	0	1	1	0	0	5
15	1	1	0	1	1	1	0	0	1	0	6
16	0	1	1	0	0	0	1	0	1	1	5
17	1	1	1	1	1	0	1	1	0	0	7
18	0	1	1	1	0	0	0	1	1	1	6
19	1	0	0	1	1	1	1	0	0	0	5
20	1	1	1	0	0	0	1	1	1	0	6
21	0	0	0	0	1	0	0	0	1	1	3
22	1	1	1	0	1	0	0	1	1	1	7
23	0	0	1	1	1	0	0	1	1	0	5
24	1	0	0	0	1	0	1	0	0	1	4
25	1	1	1	1	0	0	0	1	1	0	6
26	1	0	0	1	1	0	0	1	1	1	6
27	0	1	0	1	0	0	0	0	0	1	3

28	0	1	1	0	1	1	1	1	0	1	7
29	1	1	1	1	0	1	1	1	1	0	8
30	0	1	0	0	0	0	0	0	0	1	2
Jumlah	20	19	18	18	16	11	16	15	20	14	

Dari tabel di atas dapat dihitung tingkat kesukaran masing-masing butir tes, sebagai berikut :

$$TK_1 = \frac{20}{30} = 0{,}67 \quad \text{(TK sedang)}$$

$$TK_2 = \frac{19}{30} = 0{,}63 \quad \text{(TK sedang)}$$

$$TK_3 = \frac{18}{30} = 0{,}60 \quad \text{(TK sedang)}$$

$$TK_4 = \frac{18}{30} = 0{,}60 \quad \text{(TK sedang)}$$

$$TK_5 = \frac{16}{30} = 0{,}53 \quad \text{(TK sedang)}$$

$$TK_6 = \frac{11}{30} = 0{,}36 \quad \text{(TK sedang)}$$

$$TK_7 = \frac{16}{30} = 0{,}53 \quad \text{(TK sedang)}$$

$$TK_8 = \frac{15}{30} = 0{,}50 \quad \text{(TK sedang)}$$

$$TK_9 = \frac{20}{30} = 0{,}67 \quad \text{(TK sedang)}$$

$$TK_{10} = \frac{14}{30} = 0{,}47 \quad \text{(TK sedang)}$$

Dari hasil pengujian di atas tampak bahwa semua butir tes yang diujicobakan tersebut, telah memenuhi persyaratan sebagai butir tes yang baik, karena semua butir tes termasuk ke dalam golongan butir tes dengan TK sedang.

Meskipun demikian, harus difahami bahwa tingkat kesukaran bukanlah satu-satunya indikator dari tes yang baik. Bahkan Naga (1997) berpendapat bahwa dalam keadaan tertentu, tingkat kesukaran dapat saja diabaikan dengan mempertimbangkan indikator lain dari karakteristik butir tes, misalkan indikator validitas butir tes dan reliabilitas tes.

BAB VI
MENGUBAH SKOR MENJADI NILAI

Pada bagian sebelumnya sudah ditegaskan perbedaan antara skor dengan nilai. Skor merupakan ukuran kuantitatif berupa angka yang dihasilkan dari pengukuran, sedangkan nilai adalah adalah ukuran kualitatif yang merupakan hasil pengolahan terhadap skor. Dari perbedaan tersebut, juga terdapat dua istilah penting, yakni pengukuran dan penilaian. Mengukur adalah membandingkan ukuran obyek yang diukur terhadap standar tertentu. Pengukuran menggunakan instrumen pengukuran. Sedangkan menilai adalah merubah skor menjadi nilai. Penilaian tidak membutuhkan instrumen. Penilaian membutuhkan teknik dan kemampuan analisis dari evaluator, dan kadangkala juga membutuhkan alat bantu dan beberapa konsep perhitungan.

Terdapat dua pendekatan yang digunakan dalam merubah skor menjadi nilai. Dengan kata lain, ada dua pendekatan penilaian,, yakni Penilaian Acuan Normatif dan Penilaian Acuan Patokan (Aries, 2011, Mansyur, dkk, 2009, Cyrs, 2010). Kedua jenis pendekatan penilaian tersebut akan dibahas sebagai berikut:

A. Penilaian Acuan Normatif (PAN)

Penilaian acuan normatif atau PAN, adalah pendekatan penilaian yang diberikan kepada seorang peserta didik dibandingkan peserta didik lainnya dalam kelompoknya. Pendekatan PAN merupakan pendekatan komparatif antar peserta didik, dimana kinerja seorang peserta didik

dibandingkan dengan kinerja peserta didik yang lain dalam kelompoknya (Cyrs, 2010).

Pemberian nilai kepada seorang peserta didik dilakukan setelah evaluator mengetahui posisinya terhadap peserta didik yang lain. Standar kelulusan seseorang dalam ujian, baru dapat ditentukan setelah evaluator mengetahui hasil tes peserta didik secara keseluruhan sebagai pembanding. (Aries, 2011).

Berdasarkan pendekatan PAN, posisi seorang peserta tidak dapat dipisahkan dari posisi dirinya dan peserta didik lain dalam kelompoknya. Sehingga jika kita mengatakan bahwa seorang peserta didik adalah seseorang yang pintar, maka kepintaran yang dimaksud di sini adalah bahwa peserta didik tersebut lebih pintar dibandingkan dengan peserta didik di dalam kelompoknya. Artinya, kita tidak bisa memberikan nilai atau menyebutkan posisi seorang peserta didik, tanpa menyebutkan kelompoknya. Dalam penilaian PAN, ada ketergantungan antara nilai peserta didik dengan kelompoknya. Hasil penilaian seseorang pada kelompok A, hanya berlaku pada kelompok A tersebut, tidak dapat diberlakukan atau dibandingkan secara langsung dengan nilai peserta didik pada kelompok B.

Salah satu kelemahan pendekatan PAN, adalah kelayakan nilai antar kelompok yang tidak dapat dibandingkan secara langsung. Dalam sistem pembelajaran saat ini, banyak terdapat kelas paralel, di mana setiap kelas diajar oleh guru yang berbeda-beda. Sekalipun menggunakan bahan ajar dan materi pembelajaran yang sama, hasil pembelajaran antar kelas kurang layak untuk dibandingkan secara langsung karena berbedanya gaya dan teknik guru dalam menyampaikan pembelajaran. Misalkan untuk mata pelajaran Matematika di kelas IV-A SD, diajarkan oleh guru A. Sedangkan mata pelajaran Matematika di kelas IV-B SD diajarkan oleh guru B. Maka hasil belajar Matematika antara kelas A dan kelas B tidak dapat dibandingkan secara langsung. Seorang peserta yang memperoleh nilai B+ pada kelas A, tidak

bisa dibandingkan dengan peserta didik yang juga memperoleh nilai B+ pada kelas B. Artinya, nilai B+ pada kelas A hanya berlaku pada kelas A, dan nilai B+ pada kelas B hanya berlaku di kelas B. Nilai keduanya akan dapat dibandingkan setelah dilakukan pengolahan nilai menggunakan analisis statistika.

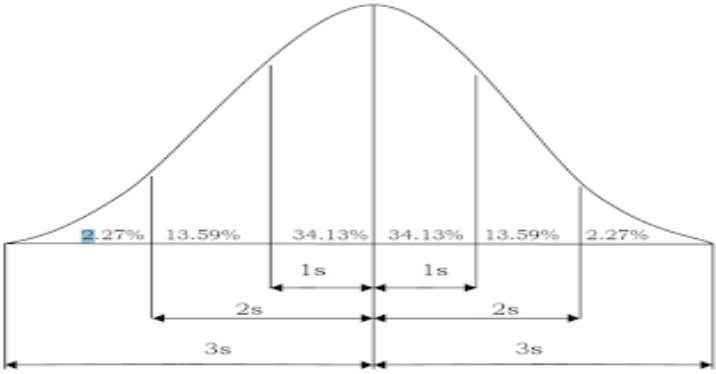
Penilaian menggunakan pendekatan PAN akan sangat berguna untuk menilai kemampuan peserta didik dalam kelompok-kelompok kecil, sehingga posisi seorang peserta didik dapat dibandingkan dengan posisi peserta didik lainnya dalam kelompok tersebut. Dalam pelaksanaan pembelajaran, suatu kelas merupakan kelompok kecil yang dapat menjadi batasan norma, sehingga guru dapat menentukan posisi seorang peserta didik di dalam kelas tersebut. Penentuan nilai seseorang dalam kelas akan sangat membantu guru untuk menentukan tindakan apa yang dapat dilakukan terhadap peserta didik tersebut, atau tindakan lainnya agar sebagian besar peserta didik bisa mencapai tujuan pembelajaran. Sebagai contoh, dari hasil penilaian menggunakan pendekatan PAN, guru dapat mengetahui peserta didik mana yang lebih pintar dibandingkan dengan yang lain, sehingga dapat ditunjuk sebagai pemandu bagi peserta didik yang kurang pintar.

Kelebihan lain dari penggunaan pendekatan PAN dalam penilaian adalah, sebaran nilai akan lebih merata dalam satu kelompok, dari nilai terendah hingga tertinggi. Penggunaan asumsi distribusi Normal, akan menyebabkan sebagian besar nilai akan berada pada posisi di tengah, dan hanya sedikit peserta tes yang memperoleh nilai rendah dan nilai tinggi.

Untuk merubah skor hasil pengukuran menjadi nilai dengan menggunakan pendekatan PAN, maka evaluator dapat menggunakan 2 pendekatan, yakni menggunakan kurva distribusi Normal, dan menggunakan pendekatan logis yang disepakati.

1. **Menggunakan kurva distribusi Normal**

Untuk merubah skor menjadi nilai menggunakan pendekatan kurva Normal, maka guru atau evaluator harus terlebih dahulu menghitung dan mengetahui rata-rata dan simpangan baku (standar deviasi) dari skor pengukuran tersebut. Secara statistika, rata-rata dan simpangan baku merupakan dasar bagi evaluator dalam perubahan skor menjadi nilai, tergantung kepada sebaran nilai yang diinginkan evaluator. Dengan berpatokan pada kurva distribusi Normal, maka guru atau evaluator dapat membagi nilai peserta didik menjadi tiga bagian, empat bagian, lima bagian, atau bahkan lebih. Dengan asumsi bahwa sebaran kurva Normal terbagi menjadi 6 daerah, yang mana setiap daerah tersebut merentang dalam satu kali simpangan baku (SB atau s), maka evaluator dapat merubah skor menjadi beberapa bagian nilai yang dikehendaki. Hal tersebut diilustrasikan sebagai berikut:



Gambar 6.1. Pembagian dalam kurva Normal

Dengan demikian, jika evaluator akan mengubah skor hasil pengukuran menjadi menjadi 6 kategori nilai, misalkan dari A, A-, B+, B, B-, dan C+, maka evaluator terlebih dahulu menghitung rata-rata (ẍ) dan simpangan baku (SB atau s). Kemudian membagi kurva Normal

menjadi 6 bagian. Karena kurva Normal sudah terdiri dari 6 bagian yang sama dalam rentang simpangan baku masing-masing, maka satu bagiannya adalah sama dengan 1 kali simpangan baku atau 1 SB. dengan demikian, pembagian interval skornya adalah sbb:

Bagian 1 : Rata-rata sampai 1 SB atau rata-rata + 1 SB

Bagian 2 : Rata-rata sampai -1 SB atau rata-rata – 1 SB

Bagian 3 : Rata-rata + 1 SB + 1 SB atau Rata-rata sampai 2 SB

Bagian 4 : Rata-rata – 1 SB + (-1 SB), atau rata-rata – 2 SB

Bagian 5 : Rata-rata + 2 SB atau lebih

Bagian 6 : Rata-rata – 2 SB ke bawah

Jika pembagian tersebut diurutkan, maka diperoleh pedoman perubahan skor menjadi nilai sebagaimana tabel berikut:

Tabel 6.1 Contoh pedoman penilaian untuk 6 kategori

Batasan rentang skor	Kategori nilai
(ẍ + 2 SB) ke atas	A
(ẍ + 1 SB) sampai (ẍ + 2 SB)	A-
(ẍ) sampai (ẍ + 1 SB)	B+
(ẍ - 1 SB) sampai (ẍ)	B
(ẍ - 2 SB) sampai (ẍ - 1 SB)	B-
(ẍ - 2 SB) ke bawah	C+

Keterangan:

ẍ = Rata-rata skor hasil pengukuran

SB = Simpangan baku atau standar deviasi skor hasil pengukuran

Dengan pendekatan yang sama, maka evaluator dapat membagi kurva Normal menjadi jumlah bagian sesuai dengan kebutuhan penilaian. Misalkan evaluator akan merubah skor menjadi nilai dalam 3 kategori, yakni BAIK, CUKUP, KURANG, maka kurva Normal dapat dibagi menjadi 3 bagian, sehingga setiap bagian terdiri atas 2 kali simpangan baku atau 2SB. Pedoman perubahan skor menjadi nilai adalah sebagai berikut:

Tabel 6.2 Contoh pedoman penilaian untuk 3 kategori

Batasan rentang skor	Kategori nilai
(ẍ + 1 SB) ke atas	BAIK
(ẍ - 1 SB) sampai (ẍ - 1 SB)	CUKUP
(ẍ - 1 SB) ke bawah	KURANG

Dengan cara yang sama, kita dapat mengubah skor menjadi nilai, dalam beberapa kategori yang kita inginkan, contohnya adalah berdasarkan tabel pedoman berikut:

Tabel 6.3 Contoh pedoman penilaian untuk 4 kategori

Batasan rentang skor	Kategori nilai
(ẍ + 1,5 SB) ke atas	Kategori 1
(ẍ) sampai (ẍ + 1,5 SB)	Kategori 2
(ẍ - 1,5 SB) sampai (ẍ)	Kategori 3
(ẍ - 1,5 SB) ke bawah	Kategori 4

Tabel 6.4 Contoh pedoman penilaian untuk 5 kategori

Batasan rentang skor	Kategori nilai
(ẍ + 1,8 SB) ke atas	Kategori 1
(ẍ + 0,6 SB) sampai (ẍ + 1,8 SB)	Kategori 2

(ẍ - 0,6 SB) sampai (ẍ + 0,6 SB)	Kategori 3
(ẍ - 1,8 SB) sampai (ẍ - 0,6 SB)	Kategori 4
(ẍ - 1,8 SB) ke bawah	Kategori 5

Sebagai contoh, berikut ini disajikan hasil tes Matematika 25 orang peserta didik kelas V SD, yakni:

Tabel 6.5. Contoh skor hasil tes Matematika

Kode PD	Skor tes (Y)
1	15
2	12
3	11
4	19
5	20
6	22
7	17
8	18
9	19
10	15
11	16
12	14
13	15
14	16
15	17
16	18
17	15

18	14
19	16
20	15
21	15
22	16
23	15
24	17
25	14
Jumlah	

Agar skor hasil tes tersebut dapat diubah menjadi nilai, maka evaluator harus menghitung rata-rata (ẍ) dan simpangan baku (SB) dari skor tersebut, yakni dengan tabel bantu analisis sebagai berikut:

Tabel 6.6. Tabel bantu menghitung rata-rata dan SB

Kode PD	Skor tes (X)	X^2
1	15	225
2	12	144
3	11	121
4	19	361
5	20	400
6	22	484
7	17	289
8	18	324
9	19	361
10	15	225
11	16	256

12	14	196
13	15	225
14	16	256
15	17	289
16	18	324
17	15	225
18	14	196
19	16	256
20	15	225
21	15	225
22	16	256
23	15	225
24	17	289
25	14	196
Jumlah	401	6573

Dari tabel di atas, dapat dihitung:

Rata-rata skor = Jumlah data/banyak data

= 401/25

= 16,04

$$\text{Simpangan Baku} = V\frac{\Sigma X^2 - (\Sigma X)^2/(N)}{(N-1)}$$

$$\text{Simpangan Baku} = V\frac{6573 - (401)^2/(25)}{(25-1)}$$

$$\text{Simpangan Baku} = V \frac{6573 - (160.801/(25)}{(24)}$$

$$\text{Simpangan Baku} = V \frac{6.573 - (6.432,04)}{(24)}$$

$$\text{Simpangan Baku} = V \frac{140,96}{(24)}$$

Simpangan Baku = V 5,873

Simpangan Baku = 2,42 (dibulatkan 2 desimal)

Dengan demikian, diperoleh rata-rata ẍ = 16,04 dan simpangan baku SB = 2,42. Berdasarkan perhitungan tersebut, jika evaluator akan merubah skor tersebut menjadi 6 kategori, yakni A, A-, B+, B, B-, dan C+, maka kategorisasinya adalah:

(ẍ + 2 SB) : 16,04 + 2 (2,42) = 16,04 + 4,84 = 20,88

(ẍ + 1 SB) : 16,04 + 1 (2,42) = 16,04 + 2,42 = 18,46

(ẍ - 1 SB) : 16,04 - 1 (2,42) = 16,04 - 2,42 = 13,62

(ẍ - 2 SB) : 16,04 - 2 (2,42) = 16,04 - 4,84 = 11,20

Dari perhitungan di atas, diperoleh tabel pedoman sebagai berikut:

Tabel 6.7. Contoh pedoman perubahan skor tes Matematika menjadi 6 kategori

Batasan rentang skor	Rentang skor	Kategori nilai
(ẍ + 2 SB) ke atas	20,88 ke atas	A
(ẍ + 1 SB) sampai (ẍ + 2 SB)	18,46 sampai 20,87	A-
(ẍ) sampai (ẍ + 1 SB)	16,04 sampai 18,45	B+
(ẍ - 1 SB) sampai (ẍ)	13,62 sampai 16,03	B
(ẍ - 2 SB) sampai (ẍ - 1 SB)	11,20 sampai 13,61	B-
(ẍ - 2 SB) ke bawah	11,20 ke bawah	C+

Berdasarkan tabel pedoman di atas, maka evaluator dapat merubah skor setiap peserta tes menjadi nilai C+ hingga nilai A sebagai berikut:

Tabel 6.8. Contoh nilai tes Matematika 6 kategori

Kode PD	Skor tes (Y)	Nilai
1	15	B
2	12	B-
3	11	C+
4	19	A-
5	20	A-
6	22	A
7	17	B+
8	18	B+
9	19	A-

10	15	B
11	16	B
12	14	B
13	15	B
14	16	B
15	17	B+
16	18	B+
17	15	B
18	14	B
19	16	B
20	15	B
21	15	B
22	16	B
23	15	B
24	17	B+
25	14	B

Sedangkan jika skor tersebut akan diubah menjadi nilai dalam 3 kategori, misalkan nilai BAIK, CUKUP, KURANG, maka tabel pedomannya menjadi:

Tabel 6.9. Contoh pedoman perubahan skor tes Matematika menjadi 3 kategori

Batasan rentang skor	Rentang skor	Kategori nilai
(ẍ + 1 SB) ke atas	18,46 ke atas	BAIK
(ẍ - 1 SB) sampai (ẍ - 1 SB)	13,63 sampai 18,45	CUKUP

(ẍ - 1 SB) ke bawah	13,62 ke bawah	KURANG

Berdasarkan tabel pedoman di atas, maka evaluator dapat merubah skor setiap peserta tes menjadi nilai KURANG hingga nilai BAIK sebagai berikut:

Tabel 6.10. Contoh nilai tes Matematika 3 kategori

Kode PD	Skor tes (Y)	Nilai
1	15	B
2	12	KURANG
3	11	KURANG
4	19	BAIK
5	20	BAIK
6	22	BAIK
7	17	CUKUP
8	18	CUKUP
9	19	BAIK
10	15	CUKUP
11	16	CUKUP
12	14	CUKUP
13	15	CUKUP
14	16	CUKUP
15	17	CUKUP
16	18	CUKUP
17	15	CUKUP
18	14	CUKUP
19	16	CUKUP
20	15	CUKUP

21	15	CUKUP
22	16	CUKUP
23	15	CUKUP
24	17	CUKUP
25	14	CUKUP

2. Menggunakan kriteria logis

Yang dimaksud dengan kriteria logis di sini adalah batasan-batasan rentang atau interval skor yang disepakati oleh beberapa fihak atau para evaluator, sehingga hal itu dapat dijadikan dasar dan pedoman dalam memberi nilai. Contoh pedoman perubahan skor menjadi nilai yang menggunakan kriteria logis adalah pedoman pemberian nilai untuk mata kuliah tertentu yang berlaku di beberapa perguruan tinggi, yakni:

Tabel 6.11. Contoh pedoman penilaian secara logis

Batasan rentang skor	Nilai angka	Nilai huruf
0-39	0	E
40-55	1	D
56-69	2	C
70-79	3	B
80-100	4	A

Dari tabel di atas tampak bahwa interval dari masing-masing kategori dan kriteria adalah berbeda-beda. Hal ini didasarkan pada kesepakatan dan pengalaman semata.

3. Membandingkan skor antar kelompok

Sebagaimana dijelaskan pada bagian sebelumnya, ciri utama penilaian menggunakan pendekatan PAN adalah, posisi atau nilai seseorang hanya dapat dibandingkan

dengan nilai orang lain dalam kelompoknya. Hal ini menimbulkan kesulitan, ketika guru atau evaluator berhadpan dengan beberapa kelas parallel. Adanya kelas kelas parallel, menimbulkan kebutuhan komparatif, yakni kebutuhan untuk membandingkan nilai antar kelompok. Karena norma atau pembandingnya adalah di dlalam kelompok, maka nilai yang diolah menggunakan pendekatan PAN tersebut tidak dapat dibandingkan secara langsung. Pembandingan nilai antar kelompok membutuhkan analisis statistika menggunakan pendekatan kurva Normal standar.

Secara statistika, suatu sebaran skor yang berdistribusi Normal dan disusun berdasarkan norma tertentu dapat dibandingkan dengan sebaran lain yang sejenis, dengan cara menyatakan sebaran tersebut ke dalam distribusi Normal standar (Glass & Hopkins, 1984). Distribusi Normal standar sering disebut sebagai distribusi Z. Caranya adalah dengan mengubah skor yang diperoleh seseorang ke dalam skor Z, dengan rumus sebagai berikut:

$$Z_{score} = (X_i - \ddot{x}\ )/SBx)$$

Dimana :

Z_{score} = nilai z seseorang

X_i = skor hasil pengukuran yang diperoleh seseorang

$\ddot{x}$ = rata-rata skor pada kelompok yang bersangkutan

SBx = simpangan baku skor pada kelompok yang bersangkutan

Misalkan diketahui skor tes mata pelajaran IPS yang diperoleh seorang peserta tes yang berada kelas A adalah 80, dan rata-rata skor hasil tes IPS pada kelas A adalah 60

dengan simpangan baku 15, maka skor Z yang diperoleh peserta didik tersebut adalah:

$$Z_{score} = (X_i - \ddot{x}) / SBx)$$
$$= (80-60)/15$$
$$= 20/15$$
$$= 1,33$$

Dengan cara merubah semua skor peserta tes menjadi skor Z, maka kita dapat membandingkan semua nilai peserta didik, baik di dalam kelompoknya maupun terhadap kelompok lainnya.

Contoh: Berikut ini adalah data hasil tes mata pelajaran PPKn kelas IV-A, IV-b, dan IV-c sebuah sekolah dasar.

Tabel 6.12. Contoh hasil tes 3 kelas paralel

Skor hasil tes PPKn kelas:					
IV-a		IV-b		IV-c	
Kode PD	Skor	Kode PD	Skor	Kode PD	Skor
A	7	O	6	AA	6
B	8	P	5	AB	7
C	6	Q	6	AC	6
D	7	R	7	AD	7
E	8	S	6	AE	8
F	6	T	7	AF	7
G	7	U	8	BB	6
H	8	V	7	BC	7
I	7	X	6	BD	6

J L M	6 7 8	Y Z	7 6	BE BF BG BJ BK	7 6 7 8 6
Jumlah	85		71		94
Rata-rata	7.08		6.45		6.71
SB	0.79		0.82		0.73

Dari tabel di atas, dapat kita ubah masing-masing skor peserta didik pada setiap kelas menjadi skor Z, dengan menggunakan rumus Z-score di atas, sehingga akan diperoleh skor Z masing-masing sebagai berikut:

Tabel 6.13. Skor Z masing-masing peserta didik pada 3 kelas parallel

Skor hasil tes PPKn kelas:					
IV-a		IV-b		IV-c	
Kode PD	Skor Z	Kode PD	Skor Z	Kode PD	Skor Z
A	-0.10	O	-0.55	AA	-0.97
B	1.16	P	-1.77	AB	0.40
C	-1.37	Q	-0.55	AC	-0.97
D	-0.10	R	0.67	AD	0.40
E	1.16	S	-0.55	AE	1.77
F	-1.37	T	0.67	AF	0.40
G	-0.10	U	1.89	BB	-0.97

H	1.16	V	0.67	BC	0.40
I	-0.10	X	-0.55	BD	-0.97
J	-1.37	Y	0.67	BE	0.40
L	-0.10	Z	-0.55	BF	-0.97
M	1.16			BG	0.40
				BJ	1.77
				BK	-0.97

Dari tabel di atas, dapat kita lihat bahwa skor peserta didik A yang berada pada kelas IV-a, yang semula memperoleh skor = 7, telah berubah menjadi skor Z = -0,10. Sedangkan peserta didik B yang memperoleh skor = 8 pada kelas IV-a memperoleh skor Z = 1,16. Demikian seterusnya. Dengan demikian, berdasarkan posisi skor Z yang masing-masing diperoleh peserta didik, kita dapat nyatakan bahwa:

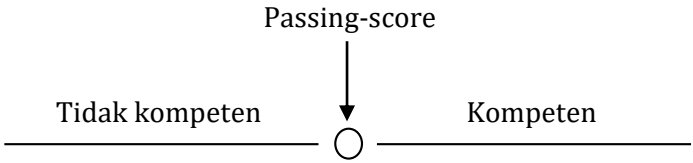
- Peserta didik A, R, dan AB sama-sama memperoleh skor 7, tetapi mereka berada pada kelas yang berbeda. Dari skor Z dapat dikatakan bahwa peserta didik yang posisi nilainya paling tinggi adalah R (skor Z = 0,67), baru disusul oleh AB (skor Z = 0,40), dan A(skor Z = -0,10).
- Peserta didik A, lebih pintar dibandingkan peserta didik O. Peserta didik O lebih pintar dari peserta didik AA.
- Peserta didik B, lebih pintar dibandingkan peserta didik AB. Peserta didik AB lebih pintar dari peserta didik P.
- Peserta didik B, lebih bodoh dibandingkan peserta didik AE. Peserta didik AE lebih bodoh dari peserta didik U.

B. Penilaian Acuan Patokan (PAP)

Penilaian acuan patokan (PAP) adalah penilaian yang didasarkan pada standar atau patokan tertentu. Perubahan skor menjadi nilai didasarkan pada angka tertentu sebagai patokan, tanpa memperhatikan posisi seorang peserta tes atau responden di dalam kelompoknya. Sebelum skor hasil tes diperoleh, evaluator sudah menetapkan kriteria sebagai standar (Aries, 2011).

Dengan demikian, norma yang digunakan dalam PAP, bukan lagi norma kelompok sebagaimana dalam PAN, tetapi yang digunakan adalah angka tertentu sebagai patokan. Angka yang dijadikan patokan tersebut sering disebut sebagai *passing-score* atau *passing-grade,* yang merupakan angka pembatas dari satu kategori dengan kategori yang lain. Dalam konteks pengukuran aspek tertentu, *passing-grade* bahkan dapat bernilai prediktif untuk memperkirakan kesuksesan peserta tes di masa dating (Cyrs, 2010).

Contoh penggunaan pendekatan PAP adalah dalam penilaian kompetensi. Dalam penilaian kompetensi, ada skor tertentu yang dijadikan sebagai standar atau patokan, sehingga patokan itu dijadikan dasar dalam menentukan nilai seseorang, apakah kompetensinya sudah tercapai atau belum. Hal tersebut dapat diilustrasikan sebagai berikut:



Penilaian yang digunakan di sekolah pada saat ini, umumnya adalah penilaian menggunakan pendekatan PAP, karena sekolah di Indonesia saat ini menggunakan kurikulum berbasis kompetensi. Konsep merdeka belajar yang diusung akhir-akhir ini, lebih berorientasi ada perubahan pendekatan proses pembelajaran, tetapi tetap mengacu pada kurikulum berbasis kompetensi. Dengan demikian, pengukuran dan

penilaiannya berorientasi pada pengukuran dan penilaian kompetensi menggunakan PAP. Dalam penilaian di sekolah, digunakan istilah Kriteria Ketuntasan Minimal (KKM), yang dijadikan sebagai patokan atau standar untuk menentukan tercapai tidaknya kompetensi yang diharapkan dalam tujuan pembelajaran. angka KKM adalah *passing-score* dalam penilaian hasil pembelajaran di sekolah. Peserta didik yang skornya sama dengan atau berada di atas KKM, disebut sebagai peserta didik yang tuntas, sedangkan peserta didik yang skornya berada di bawah KKM disebut sebagai peserta didik yang tidak tuntas.

Pada penilaian menggunakan PAP, sering hanya terdapat 2 kategori nilai yang bersifat dikotomik, misalkan kompeten-tidak kompeten, lulus-tidak lulus, atau tuntas-tidak tuntas. Risiko dari nilai dikotomik ini antara lain, sebaran nilai peserta tes tidak menyebar merata. Adakalanya nilai mengumpul di atas kriteria, atau bisa pula di atas kriteria, sehingga variasi kemampuan peserta tes tidak tergambar dengan baik dan merata. Tetapi kelebihnnya adalah, evaluator memiliki standar yang sama dalam memberikan penilaian, sehingga relatif nilai seseorang dalam suatu kelompok dapat dibandingkan dengana nilai orang lain pada kelompoknya, bahkan dengan nilai orang lain di luar kelompoknya yang menggunakan standar yang sama.

Fokus utama penilaian menggunakan PAP adalah pada *passing-score*nya. Angka pada *passing-score* merupakan patokan yang membagi peserta tes menjadi dua bagian, antara yang lulus dengan yang tidak, yang tuntas dengan yang tidak tuntas. Dengan adanya *passing-score* sebagai patokan, maka evaluator tinggal merubah skor yang diperoleh peserta tes menjadi 2 kelompok nilai berdasarkan *passing-score* yang telah ditentukan.

Terdapat beberapa cara yang umumnya digunakan untuk menetapkan *passing-score* atau *passing-grade* dalam penilaian PAP, antara lain:

1. Berdasarkan Kesepakatan

Dalam konteks ini, *passing-score* ditentukan oleh sekelompok evaluator berdasarkan pengalaman dan kesepakatan. Contoh penggunaan cara ini adalah kesepakatan yang digunakan guru dalam menentukan Kriteria Ketuntasan Minimal (KKM) pada suatu ata pelajaran tertentu. Dalam suatu wilayah tertentu, biasanya terdapat MGMP (Musyawarah Guru Mata Pelajaran), yang merupakan wadah guru yang mengajar pada mata pelajaran yang sama. Kelompok MGMP ini juga bisa dinggap sebagai kelompok evaluator untuk mata pelajaran tertentu. Dalam musyawarah MGMP mata pelajaran Matematika kelas IV SD misalnya, para guru menentukan KKM untuk mata pelajaran Matematika di satu wilayah tertentu adalah 7,5. Angka KKM=7,5 adalah merupakan *passing-score* yang disepakati sebagai patokan.

2. Berdasarkan Perhitungan

Kadangkala evaluator perlu menentukan angka tertentu yang eksak sebagai *passing-score* dalam penilaian kompetensi. Untuk itu diperlukan amalisis berdasarkan pendekatan, metode, dan perhitungan tertentu.

Secara umum terdapat 3 jenis pendekatan yang digunakan dalam penetapan *passing-score.* Pertama adalah pendekatan yang berpusat pada tes yang digunakan (*test centered models).* Kedua, pendekatan yang berpusat pada peserta tes (*examined centered models),* dan yang ketiga adalah pendekatan kompromistis antara keduanya (Bulkani, 1999). Pada pendekatan pertama, penetapan *passing-score* dimulai dengan memperkirakan skor minimal yang harus diperoleh oleh peserta tes untuk mencapai kompetensi tertentu menggunakan tes yang dicari *passing-score*nya. Tampak di sini bahwa pusat perhatian evaluator adalah pada instrumen tes yang digunakan dan skor tes yang

dihasilkan. Pada pendekatan kedua, evaluator menetapkan *passing-score* dengan cara membandingkan hasil pengukurannya dengan hasil pengukuran sejenis yang telah dilaksanakan terdahulu, baik terhadap peserta tes yang sama maupun pada peserta tes dari kelompok berbeda. Tampak bahwa pendekatan kedua ini berpusat pada peserta tes atau responden pengukuran. Sedangkan pendekatan ketiga, menggunakan pendekatan gabungan antara keduanya sehingga disebut pendekatan kompromistis.

Beberapa contoh metode yang digunakan dalam menetapkan *passing-score* secara eksak adalah sebagai berikut:

a. Metode Ebel

   Metode Ebel menggunakan pendekatan *test centered models.* Metode ini memerlukan beberapa orang evaluator ahli atau *rater*, yang dianggap pakar dalam bidang kompetensi yang diukur dan akan ditetapkan *passing-score*nya.

   Untuk menentukan *passing-score* menggunakan metode ini, para evaluator ahli (sering pula disebut penilai partisipan), diminta untuk menelaah butir tes sambil melakukan 3 hal, yakni:

   - Memperkirakan tingkat kesukaran masing-masing butir tes. Umumnya butir tes tersebut dikategorikan ke dalam butir mudah-sedang-sulit.
   - Memperkirakan dan menelaah relevansi isi tes (validitas isi tes) terhadap kompetensi yang akan diukur. Relevansi tersebut umumnya dikategorikan menjadi sangat penting-cukup penting-dapat digunakan-diragukan.
   - Memperkirakan persentase jawaban benar, atau menentukan kemungkinan persentase jawaban

benar peserta tes berdasarkan tingkat kesukaran butir tes dan relevansi isi tes.

Sebagai contoh, tabel berikut ini adalah hasil analisis oleh 3 orang evaluator ahli terhadap 30 butir tes kompetensi.

Tabel 6.14. Contoh penentuan *passing-score* dengan metode Ebel.

Kategori butir	% minimal jawaban benar yang diharapkan (A)	Hasil penilaian terhadap banyak butir pada kategori tertentu oleh 3 evaluator (B)	A x B
Sangat Penting			
Mudah	100	20	2.000
Sedang	100	10	1.000
Sulit	100	2	200
Jumlah		32	3.200
Cukup Penting			
Mudah	90	21	1.890
Sedang	70	7	490
Sulit	50	2	100
Jumlah		30	2.480
Dapat			

Diterima Mudah Sedang Sulit	80 60 40	8 5 2	640 300 80
Jumlah		15	1.020
Diragukan Mudah Sedang Sulit	 70 50 30	 7 4 2	 490 200 60
Jumlah		13	750
Jumlah semua		3x30=90	7.450

Dari tabel di atas dapat ditetapkan persentase kritis sebagai dasar perhitungan *passing-score*, yakni : 7.450/90 = 82,78 %. Ini berarti bahwa paling tidak peserta tes harus benar menjawab sebanyak 82,78% agar dapat dikategorikan mencapai kompetensi yang diukur. Untuk 30 butir soal tes, berarti paling tidak peserta tes harus menjawab benar sebanyak 82,78% x 30 butir = 24,83 butir (dibulatkan menjadi 25 butir). Ini berarti bahwa, seorang peserta tes dikatakan memiliki kompetensi yang diinginkan, jika minimal mampu menjawab dengan benar sebanyak 25 butir dari 30 butir tes kompetensi yang diberikan.

Metode Ebel dapat dikembangkan dan digunakan pada tes kompetensi berbentuk pilihan ganda maupun uraian. Hal ini merupakan kelebihan metode Ebel. Sedangkan kelemahannya adalah bahwa metode ini masih melibatkan butir-butir tes yang kurang baik dan diragukan. Masuknya butir-butir tes yang meragukan sebagai alat ukur, akan

menurunkan derajat validitas butir dan reliabilitas tes, karena ada kemungkinan peserta tes justeru menjawab benar pada butir-butir tes yang meragukan tersebut.

b. Metode Angoff

Metode Angoff menggunakan pendekatan berbasis tes atau *test centered models* (Angoff, 1971). Penggunaan metode ini juga memerlukan kehadiran partisipan ahli sebagai *rater* atau evaluator, yang melakukan kajian terhadap butir-butir tes kompetensi.

Pada dasarnya, terdapat kemiripan metode Angoff dengan metode Ebel, karena sama-sama mengestimasi kemampuan peserta tes dalam menjawab benar suatu tes. Pada penggunaan metode Angoff, sekelompok *rater* diminta membayangkan seandainya terdapat 100 orang peserta tes yang memiliki kompetensi minimal, kemudian diminta memperkirakan banyaknya peserta tes yang menjawab benar tes kompetensi yang dianalisis. Hal ini diulang beberapa kali, baik menggunakan partisipan ahli yang sama maupun berbeda. Hasil analisis tersebut kemudian rata-ratanya sebagai dasar untuk menetapkan *passing-score.*

Berikut ini diberikan contoh penggunaan metode Angoff, untuk menentukan *passing-score* pada 6 butir tes kompetensi, dengan 3 kali pengulangan dan 5 orang partisipan ahli. Hasil analisis diilustrasikan pada tabel berikut:

Tabel 6.15. Contoh penentuan *passing-score* dengan metode Angoff

Butir	Pengulangan	1	2	3	4	5	Jumlah	Rata-rata
1	1	45	50	60	45	55	255	51.00
	2	45	52	62	46	56	261	52.20
	3	44	50	61	47	55	257	51.40
2	1	70	75	70	75	70	360	72.00
	2	75	70	69	74	68	356	71.20
	3	74	72	70	74	70	360	72.00
3	1	80	82	85	82	85	414	82.80
	2	85	80	80	84	83	412	82.40
	3	80	82	82	85	85	414	82.80
4	1	50	52	55	60	57	274	54.80
	2	52	55	56	62	60	285	57.00
	3	55	56	58	60	58	287	57.40
5	1	100	90	95	97	100	482	96.40
	2	97	89	97	98	99	480	96.00
	3	95	97	95	96	98	481	96.20
6	1	40	45	46	44	45	220	44.00
	2	45	44	45	45	40	219	43.80
	3	45	45	45	46	43	224	44.80
Rata-rata	1							66.83
	2							67.10
	3							67.43
Rata-rata setiap putaran								67.12

Dari tabel di atas, ditemukan rata-rata estimasi yang dapat digunakan sebagai *passing-score,* dalam rentang jawaban benar antara 66,83% sampai 67,43%. Dapat pula rata-rata untuk keseluruhan putaran, yakni 67,12% yang dijadikan sebagai

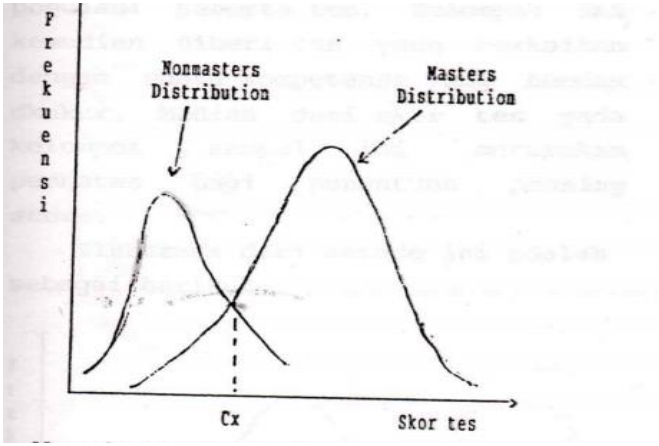
passing-score. Artinya, jika tes tersebut terdiri dari 6 butir, maka untuk dikatakan kompeten paling tidak peserta tes harus dapat menjawab dengan benar sebanyak 67,12% x 6 butir = 4,02 butir (dibulatkan menjadi 4 butir). Selanjutnya, ukuran kompeten tidaknya peserta tes ditentukan berdasarkan jawaban benar ketika menjawab tes tersebut. Peserta tes yang mampu menjawab benar minimal 4 butir tes, dikategorikan sebagai peserta tes yang kompeten.

c. Metode Kontras Grup

Metode kontras grup merupakan metode yang menggunakan pendekatan *examined centered models,* yakni dengan melihat respon dua kelompok berbeda terhadap tes kompetensi yang diberikan. Dengan demikian, metode kontras grup meerupakan metode penentuan *passing-score* berdasarkan hasil ujicoba secara empiris.

Pada penggunaan metode ini, dibutuhkan 2 grup peserta ujicoba tes yang setara, yang diambil dari populasi yang sama. Grup pertama diberikan perlakuan berupa pembelajaran tentang materi kompetensi yang akan diujicobakan, sedangkan grup kedua tidak diberikan perlakuan tersebut. Kedua grup diberikan tes kompetensi yang sama, sehingga menghasilkan dua distribusi skor hasil tes kompetensi. Distribusi skor kelompok pertama, disebut *masters distribution* atau distribusi kelompok yang menguasai kompetensi. Sebaliknya distribusi skor hasil tes grup kedua, disebut *nonmaters distribution.* Distribusi skor hasil tes kompetensi kedua kelompok ini kemudian digambarkan dalam kurva distribusi Normal pada lembar yang sama. Titik potong kedua kurva merupakan titik yang ditetapkan sebagai *passing-score.*

Metode kontras grup dapat diilustrasikan sebagai berikut :

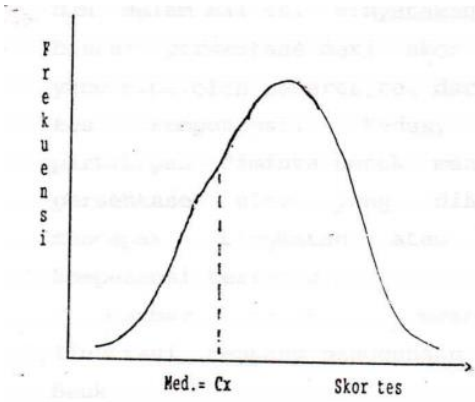


Gambar 6.2. Ilustrasi metode kontras grup

d. Metode Garis Pembatas

Metode garis pembatas juga merupakan metode yang menggunakan pendekatan *examined centered models.* Perbedaannya dengan metode kontras, pada metode ini hanya dibutuhkan satu grup ujiocoba yang telah dikenal karakteristiknya oleh para partisipan ahli, khususnya tentang kompetensi mereka. Pada awalnya, para partisipan diminta untuk menetapkan grup sebagai sampel yang diambil dari populasi peserta tes. Grup ini kemudian diberi tes kompetensi yang akan digunakan, sehingga diperoleh sebaran skor hasil tes kompetensi pada grup ini. Median dari skor hasil tes, kemudian ditetapkan sebagai *passing-score.*

Ilustrasi dari metode ini adalah sebagai berikut:



Gambar 6.3. Ilustrasi metode garis pembatas

e. Metode Beuk

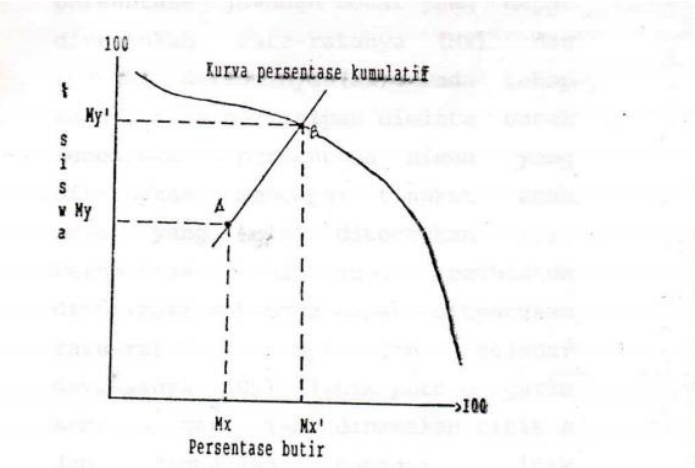
Metode Beuk merupakan metode penetapan *passing-score* menggunakan pendekatan yang kompromistis, yang menggabungkan antara *test centered models* dengan *examineed centered models* (Beuk, 1984, Fernandez, 1984). Pendekatan ini dianggap lebih cocok, karena memenuhi asumsi teori skor klasik. Dari teori tes klasik diketahui bahwa ada keterikatan dan saling ketergantungan kejadian antara karakteristik butir tes dengan karakteristik peserta tes (Naga, 1997).

Langkah-langkah penggunaan metode Beuk adalah:

- Para partisipan ahli atau *rater* diminta untuk menganalisis dan memperkirakan tingkat kompetensi minimal yang dibutuhkan peserta tes untuk dinyatakan sebagai orang yang kompeten dibidang tertentu. Perkiraan ini dinyatakan dalam bentuk persentase jawaban benar dari skor mentah yang diperoleh peserta tes kompetensi

- Para partispan ahli diminta untuk menentukan persentase peserta tes kompetensi yang diharapkan mampu mencapai tingkatan atau level kompetensi yang diharapkan.

Berikut ini diberikan ilustrasi tentang cara penentuan *passing-score* menggunakan metode Beuk.



Gambar 6.4. Ilustrasi metode Beuk

Dari gambar di atas, dapat dijelaskan bahwa pada tahap pertama para partisipan ahli diminta untuk menentukan persentase jawaban benar yang diharapkan dari para peserta tes agar mereka dinyatakan sebagai orang yang kompeten. Persentase jawaban benar ini dilambangkan dengan X. Karena perkiraan jawaban benar tersebut diberikan oleh banyak partisipan ahli, maka hasil analisis itu akan berbentuk distribusi tertentu yang dapat dihitung rata-rata atau meannya (Mx) dan simpangan bakunya (Sx). Pada tahap kedua, para partisipan ahli diminta untuk menentukan persentase peserta tes yang diharapkan mencapai

tingkat atau level kompetensi yang ditentukan. Persentase ini dilambangkan dengan Y dengan distribusi tertentu yang memiliki rata-rata atau mean My dan simpangan baku Sy. Titik Mx dilukiskan pada sumbu X dan titik My dilukis pada sumbu Y dalam sumbu koordinat Cartesius. Titik potong antara garis lurus X=Mx dan garis lurus Y=My disebut titik A yang digunakan sebagai titik referensi awal. Dari titik A dapat ditarik garis lurus dengan kemiringan atau gradien Sy/Sx sedemikian rupa sehingga memotong kurva persentase kumulatif pada titik B. Kurva persentase kumulatif dapat digambarkan dengan memperhatikan persentase kumulatif menurun. Absis atau nilai X pada titik B, yakni Mx', merupakan *passing-score.* Mx' menyatakan persentase jawaban benar pada peserta tes agar peserta tes dinyatakan memiliki kompetensi tertentu. Banyak butir tes minimal yang harus dijawab peserta tes untuk mencapai kompetensi tertentu adalah hasil perkalian antara Mx' dengan jumlah butir tes.

f. Metode Hofstee

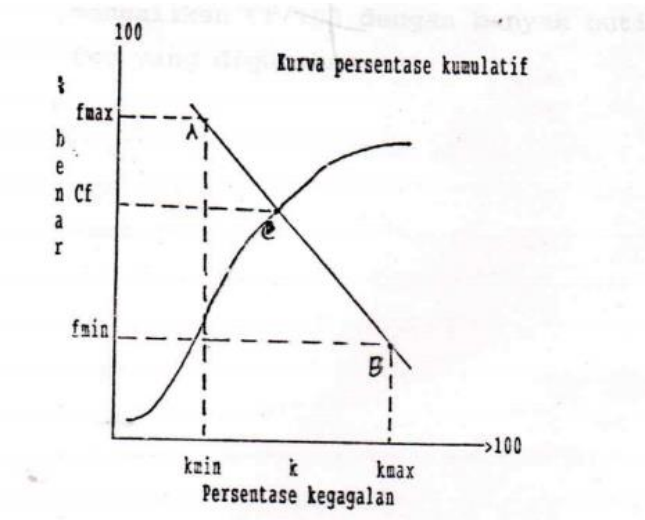
Metode Hofstee juga menggunakan pendekatan komprimistis. Metode ini digunakan dengan cara memberi 4 pertanyaan kepada para partisipan, yakni:

- Berapa skor terendah yang masih bisa diterima, terutama jika setiap peserta tes memperoleh skor tersebut pada saat pertama kali mengikuti tes?. Skor jawaban para partisipan ahli kemudian ditentukan rata-ratanya, dan dilambangkan dengan f_{min}.
- Berapa skor terendah yang masih bisa diterima, terutama jika tidak satupun peserta tes memperoleh skor tersebut pada saat pertama

kali mengerjakan tes tersebut?. Skor jawaban para partisipan ahli kemudian ditentukan rata-ratanya, dan dilambangkan dengan f_{max}.

- Berapa rata-rata kegagalan yang bisa ditolerir?. Rata-rata kegagalan ini dilambangkan dengan k_{max}.
- Berapa rata-rata kegagalan minimum yang masih bisa diterima? Rata-rata jawaban partisipan ahli ini dilambangkan dengan k_{min}.

Keempat titik tersebut kemudian dijadikan sebagai patokan untuk melukis titik koordinat pada sumbu Cartesius, yakni titik A(k_{min},f_{max}) dan titik B(k_{max}, f_{min}). Ilustrasinya adalah sebagai berikut:



Gambar 6.5. Ilustrasi metode Hofstee

Titik A dan B kemudian dihubungkan menggunakan suatu garis lurus sehingga memotong kurva persentase kumulatif menaik pada titik C. Ordinat titik C, yakni Cf, merupakan titik yang

direkomendasikan sebagai *passing-score.* Jumlah butir minimal yang harus dijawab benar oleh peserta tes agar dinyatakan mencapai kompetensi tertentu, diperoleh dengan cara mengalikan Cf/100 x jumlah butir tes.

DAFTAR PUSTAKA

Anderson, L.W., Krathwohl, D.R. (2001). A Taxonomy for Learning, Teaching, and Assesing: A Revision of Bloom's Taxonomy of Educatioanl Objectives. New York: Addison Wesley Longman, Inc.

Angoff, W.H. (1971). *Scales, Norma, and Eqivalent Score,* in R.L. Thorndike (ed). *Educational Measurement.* Washington DC: American Council on Education.

Anderson, L. W. (2010). Kerangka Landasan untuk Pembelajaran, Pengajaran, dan Asesmen Revisi Taksonomi. Yogyakarta: Pustaka Pelajar.

Aries, E.F. (2011). *Asesmen dan Evaluasi.* Malang: Aditya Media Publishing.

Arifin, Z. (2012). *Evaluasi Pembelajaran.* Jakarta: Subdit Kelembagaan Direktorat Pendidikan Tinggi Islam Kemenag RI.

Asdam, M. (2007). Pengaruh Pemberian Evaluasi Ulangan Harian terhadap Peningkatan Motivasi Belajar Bahasa Indonesia pada Siswa Tingkat SMP Kabupaten Maros. *Jurnal Pendidikan dan Kebudayaan,* Vol. 066 (13). Avalaible at: https://jurnaldikbud.kemdikbud.go.id/index.php/jpnk/article/view/360/242

Asrul, Ananda, R., dan Rosnita. (2015). *Evaluasi Pembelajaran,* Bandung: Citapustaka Media.

Beuk, C.H. (1984). A Method for Reaching Compromise between Absolute and Relative Stabdart in Examination*. Journal of Educational Measurement.* Vol. 21, pp. 147-152

Bulkani, (2020), *Index and Coefficient of Reliability on Educational Onlie Test by Repeated Measurement,* in:

Society 5.0 Fostering Spirituality and Humanity, OASE Volume 3 (1), 2020.

----------, (1999), *Penetapan Passing Score pada Penilaian Kompetensi,* Palangka Raya: FKIP Universitas Muhammadiyah Palangkaraya.

Callahan, M., & Logan, M.M. (2021). *How Do I Create Tests for My Student?* Texas: Teaching Learning, & Professional Development Centre, avalaible at: https://www.depts.ttu.edu/tlpdc/Resources/Teaching_resources/TLPDC_teaching_resources/createtests.php

Cyrs, T.E. (2010). Essential Skills for College Teaching: Chapter 19, Constructing Valid Test to Match Yours Learning Objectives. New Meksiko: New Meksiko State University.

Cronbach, L.J. (1985), *Essential of Psychological Testing,* New York: Harper and Ross

Fernandez, H.J.X. (1984). *Testing and Measurement.* Jakarta: National Education Planning, Evaluation and Curriculum Develepment.

Glass, G.V., & Hopkins, K.D. (1984), *Statistical Methods in Educational and Psychology,* New Jersey: Prentice-Hall Inc.

Kirkpatrick, D.L. (1998). *Evaluating Training Program, The Four Level. 2nd edition.* San Fransisco: Berret-Kohler Publisher Inc.

Leenknecht, M., Wijnia, L., Kohlen, M., Fryer, L., Rikers, R., & Loyen, S. (2020). Formative Assessment as Practice: The Role of Students' Motivation. *Assesment & Evaluation Journal.* Avalaible at: https://doi.org/10.1080/02602938.2020.1765228

Lestari, U.P., & Setiawan, D.F. (2017). Data Collection Methods on Learning Outcome, Student Achievement, and Academic Achievement. *Guidena Journal,* Vo. 7 (2). pp.

164-173. ISSN: 2088-9623, e-ISSN: 2442-7802 DOI: https://10.24127/gdn.v7i2.1016

Mansyur, Rasyid, H., dan Suratno. (2009). *Asesmen Pembelajaran di Sekolah.*Yogyakarta: Multi Presindo.

Mardapi, D. (2012), Pengukuran, Penilaian, & Evaluasi Pendidikan, Yogyakarta: Nuha Medika.

Mulatsih, B., (2021), Penerapan Taksonomi Bloom Revisi pada Pengembangan Soal Kimia Ranah Pengetahuan. *Ideguru: Jurnal Karya Ilmiah Guru.* Vol 6 (1). pp. 1-10. p-ISSN 2527-5712; e-ISSN 2722-2195. DOI: https://doi.org/10.51169/ideguru.v6i1.158

Naga, D.S. (2013), *Teori Sekor pada Pengukuran Mental*, Jakarta: Nagarani Citrayasa.

----------, (1992), Pengantar Teori Sekor pada Pengukuran Pendidikan, Jakarta: Besbats

Pauji, R., Trisna, B.N., dan Atsnan, M.F. (2016). Pemanfaatan Hasil Evaluasi Pembelajaran Matematika SMA di Kota Banjarmasin. *Math Didactic: Jurnal Pendidikan Matematika STIKIP PGRI Banjarmasin.* Vo. 2 (3). ISSN 2442-3041. Avalaible at: https://media.neliti.com/media/publications/176859-ID-pemanfaatan-hasil-evaluasi-pembelajaran.pdf

Ratnawulan, E., & Rusdiana, A. (2014). *Evaluasi Pembelajaran,* Bandung: Pustaka Setia.

Sax, G. (1980). Principles of Educational and Psychological Measurement and Evaluation, San Francisco: Phoenix Publishing Services Inc.

Selegi, S.F.. (2017). Pengaruh Model Evaluasi Formatif Sumatif terhadap Motivasi Belajar Mahasiswa pada Mata Kuliah Perencanaan Pengajaran Geografi. *ProsidingSeminar Nasional Program Pasca Sarjana Universitas PGRI Palembang.* November 2017.

Stufflebeam, D.L., & Shinkfield, A.J. (2007) *Evaluation Theory, Models, & Applications,* USA : John Wiley & Sons, Inc.

Widoyoko, S.E.P. (2011). Evaluasi Program Pembelajaran, Panduan Praktis bagi Pendidik dan Calon Pendidik, Yogyakarta: Pustaka Pelajar

e-mail: susantifaipriselegi@gmail.com

EVALUASI
- PEMBELAJARAN -

Dalam buku ini dibahas tentang makna evaluasi, yang di dalamnya mengandung makna pengukuran dan penilaian dengan pendekatann yang umum digunakan dalam mengukur dan menilai, serta bagaimana caranya merubah skor menjadi nilai. Selain itu, juga membahas tentang pentingnya menyusun instrumen pengukuran yang handal, dan bagaimana cara memperolehnya. Pendekatan yang digunakan dalam pengukuran hasil pembelajaran pada buku ini, menggunakan teori skor klasik supaya penggunaannya lebih praktis bagi para evaluator pemula.

BULKANI ✦ EVALUASI PEMBELAJARAN

Akademia Pustaka
Perum. BWW Madani Kavling 16, Tulungagung
https://akademiapustaka.com/
redaksi.akademia.pustaka@gmail.com
@redaksi.akademia.pustaka
@akademiapustaka
081216178398

ISBN 978-623-6264-31-0