ISSN: 2540-9417

KESAHIHAN DALAM PENYUSUNAN TES BAHASA ARAB DI MADRASAH/SEKOLAH

MOH. AININ
Universitas Negeri Malang

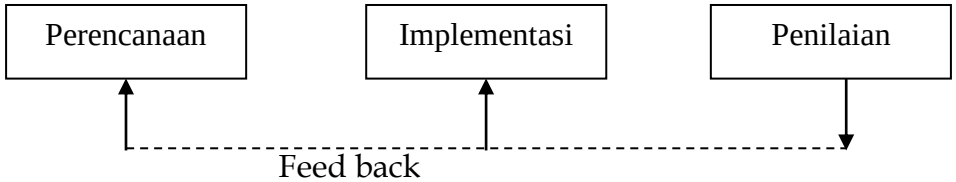
Abstrak: Penilaian merupakan bagian integral dalam sistem pembelajaran, khususnya pembelajaran bahasa Arab. Penilaian bukan saja berfungsi untuk memberikan informasi tentang keberhasilan atau kekuraangan proses dan hasil belajar, tetapi juga sebagai *feed back* untuk perbaikan sistem pembelajaran. Bahkan ia sebagai bahan refleksi terhadap kualitas tes itu sediri. Salah satu piranti dalam penilaian adalah tes. Tes yang baik adalah tes yang memiliki daya validitas (sahih). Untuk mewujudkan tes yang sahih diperlukan langkah-langkah khusus yang sistematis. Secara reduktif, langkahyang dimaksud meliputi telaah kurikulum, buku ajar, dan pembuatan kisi-kisi tes.

Kata kunci: kesahihan, penyusunan tes, kisi-kisi tes.

A. PENDAHULUAN

Evalusi atau penilaian merupakan salah satu bagian integral dalam suatu pembelajaran yang dilaksanakan oleh guru. Sebagai bagian integral, evaluasi memiliki peranan yang signifikan untuk menentukan keberhasilan pembelajaran itu sendiri, baik dari sisi proses maupun dari sisi hasil. Bahkan evaluasi sendiri tidak hanya terkait dengan pemberian keputusan terhadap proses dan hasil belajar, tetapi juga terkait dengan keberadaan dirinya sebagai umpan balik (*feed back*) untuk perbaikan sistem pembelajaran.Bahkan penilaian juga dapat digunakan sebagai refleksi terhadap kualitas alat penilaian itu sendiri atau menilai sebuah tes (*testing the test*) (Mc Namara, 2008).

Terkait dengan tugas guru, Copper (1979) menegaskan bahwa tugas utama pendidik profesional sebagai pengambil keputusan dalam proses belajar-mengajar meliputi perencanaan, implementasi, dan penilaian. Menurut Cooper (1977), pada tahap perencanaan, tugas operasional pendidik meliputi: analisis kebutuhan siswa, merumuskan tujuan yang sesuai, menentukan model dan strategi pembelajaran untuk mencapai tujuan, dan merencanakan bahan ajar. Pada tahap implementasi, pendidik melaksanakan apa yang telah direncanakan. Sementara itu, pada tahap evaluasi, pendidik menilai sejauhmanakah pembelajaran tercapai, baik yang terkait dengan proses maupun hasil belajar. Ketiga tugas utama pendidik tersebut oleh Cooper (1977) dibagankan sebagai berikut.



ISSN: 2540-9417

Secara lebih spesifik, fungsi penilaian dapat dikemukakan sebagai berikut. (a) untuk mengukur hasil belajar siswa, (b) untuk mengetahui tingkat capaian tujuan pembelajaran, (c) untuk mengetahui tingkat keberhasilan pendidik dalam pembelajaran, (d) untuk menentukan materi ajar dan kompetensi yang harus dipelajari, (e) untuk mengetahui strategi pembelajaran yang lebih tepat, (f) untuk penentuan kenaikan kelas siswa, (g) sebagai informasi bagi orang tua siswa (wali murid) tentang tingkat kemampuan putranya, (h) untuk pengelompokan siswa berdasarkan kemampuan, (i) untuk meningkatkan motivasi belajar siswa, (y) untuk mengetahui kekurangan dan kelebihan proses pembelajaran, dan (k) untuk mendiagnosis kesulitan belajar siswa (Al-Khuli, 1986 dan Asrori, et al., 2012). Bahkan penilaian juga dapat digunakan sebagai refleksi terhadap kualitas alat penilaian itu sendiri atau menilai sebuah tes (*testing the test*) (Mc Namara, 2008).

Penilaian sebagai bagian integral dalam pembelajaran, khususnya Pembelajaran Bahasa Arab (PBA) cenderung dipandang dan diaplikasikan secara parsial. Penilaian lebih diposisikan sebagai kegiatan untuk mengumpulkan informasi yang terkait dengan hasil belajar peserta didik yang bersifat numerik. Penilaian hanya diidentikan dengan UN. Penilaian dipersepsi hanya sebagai alat untuk mengukur kemampuan kognitif peserta didik. Dari aspek penyelenggaraannya, penilaian hanya dilakukan pada tengah semester dan akhir semester (diidentikkan dengan tes formati da sumatif). Demikian pula, dari sisi cara dan kompetensi bahasa Arab yang diukur, penilaian lebih diposisikan bagaimana peserta didik dapat menjawab soal-soal dalam bentuk tulis (objektif atau subjektif/esai) yang kadang-kadang kualitas butir-buti soalnya kurang memenuhi persyaratan tes yang baik.

Dalam pelaksanaan penyusunan tes bahasa Arab, tidak jarang dijumpai proses penyusunannya yang "asal buat", tanpa memperhatikan prosedur yang benar. Penyusun tes membiasakan diri "mencomot" butir-butir soal dari "sana-sini" kemudian merangkaikannya menjadi sebuat tes bahasa Arab, tanpa melihat apakah butir-butir soal yang dikembangkan itu benar-benar mengukur kemampuan atau kompetensi yang seharusnya diukur. Belum lagi instrumen dan cara penilaiannya monoton. Model penyusunan tes seperti ini berdampak pada ketidaksahinan tes yang dibuat. Implikasinya instrumen tes yang dibuat kurang atau tidak memenuhi persyaratan tes yang baik.

Secara yuridis, penilaian seperti ini mengabaikan permendiknas No. 20 Tahun 2007 tentang Standar Penilaian Pendidikan. Pada butir C yang terkait dengan Teknik dan Instrumen Penilaian dikemukakan, bahwa penilaian hasil belajar menggunakan berbagai teknik. Misalnya berupa tes, observasi, penugasan perseorangan atau kelompok, dan bentuk lain yang sesuai dengan karakteristik kompetensi dan tingkat perkembangan peserta didik. Permendiknas ini mengisaratkan bahwa tes yang digunakan itu harus benar-benar mengukur kompetensi peserta didik secara sistemik, komprehensif dan mencakup multi ranah, baik ranah kognitif, psikomotorik, maupun ranah afektif. Model peniliaian inilah yang lebih dikenal dengan penilaian senyatanya atau penilaian otentik.

Untuk menilai kemampuan berbahasa Arab yang senyatanya, penilaian selayaknya berbasis performansi. Dalam penilaian performansi, keterampilan berbahasa yang diases adalah aktivitas berkomunikasi, baik lisan maupun tulis.

ISSN: 2540-9417

Pada umumnya, penilaian dalam bentuk tes performansi ini berupa kemampuan berbicara dan menulis (McNamara, 2008) atau yang lazim disebut kemampuan produktif (*al-maharah al-istintajiyah*). Meskipun demikian, penilaian terhadap kemampuan reseptif (menyimak dan membaca) tidak boleh diabaikan.

Terkait dengan permasalahan di atas, makalah ini memerikan proses penyusunan tes bahasa Arab yang sahih. Perihal yang dibahas meliputi (a) karakteristik tes yang baik, (b) penilaian dalam Kurikulum 13, dan (c) menuju tes yang sahih.

B. KARAKTERISTIK TES BAHASA YANG BAIK

Penyusunan tes bahasa, khususnya tes bahasa Arab seharusnya memperhatikan standar tes yang baik. Standar tes yang baik itu meliputi standar validitas atau kesahihan (*Ash-shidqu*), reliabilitas (*Ats-tsabat*), dan memiliki tingkat kesulitan (*darajatush shu'ubah*) dan daya beda (*darajatut tamyiz*) yang baik. Dari keempat standar ini, yang lebih prinsip dan lebih aplikatif pelaksanaannya bagi pendidik adalah standar validitas. Hal ini karena standar kesahihan ini (kesahihan logis) lebih pada proses penyelarasan antara tujuan tes dengan substansi butir-butir soal yang disusun dan untuk menentukan tingkat kesahihan ini dapat dilaksanakan sebelum tes diberlakukan. Sementara itu, ketiga standar yang terkahir ini pelaksanaannya setelah tes diberlakukan dan masih diperlukan tahapan-tahapan berikutnya.

Kesahihan atau validitas sebuah tes terkait dengan apakah butir-butir tes yang disusun mengukur yang seharusnya diukur. Gunning (1998) memberikan batasan kesahihan sebagai berikut "*validity means that a device measure what it says itu measured*". Pendapat yang sama juga dikemukakan oleh Duwaidari (2000), tes dikatakan sahih manakala tes itu mengukur sesuai dengan yang seharusnya diukur. Dalam tes bahasa arab misalnya, apabila tujuan tesnya ingin mengukur kemampuan membaca, maka tes yang disusun benar-benar mengukur kemampuan membaca, apakah membaca keras atau dan memahami isi bacaan dengan berbagai indikator yang ditetapkan. Demikian pula, apabila tujuan tes itu mengukur *maharah kalam*, maka tes yang disusun dan dilaksanakan adalah menuntut teste memformansikan *maharah kalam*, baik dalam bentuk monolog maupun dialog, bukan menulis dan membaca. Al-khuli (2000) juga memberikan gambaran bahwa tes *imla'* harus mengukur hanya kemampuan *imla'*, tes *qawa'id* harus mengukur kemampuan hanya kemampuan *qawa'id*, demikian pula tes kosa kata juga harus hanya mengukur kemampuan kosa kata. Apabila tujuan tes itu lebih dari satu, maka tes yang dusun itu juga mengacu ke tujuan tersebut, tanpa mengabaikan salah satu dari tujuan tes.

Keterpecayaan atau reliabilitas dapat dimaknai sebagai suatu keajegan tes. Artinya suatu tes dikatakan reliabel manakala tes itu diberlakukan pada kelompok yang sama secara berulang dengan menghasilkan skor yang relatif sama. Gunning (1998) mendefinisikannya "*A reliable device is one that yield consistent resuluts*". Menurut Saukah (1994), istilah reliabilitas berkaitan dengan istilah keajegan (*consistency*), keterandalan (*dependability*), kestabilan (*stability*), ketepatan (*accuracy*), dan keterdugaan (*predictibility*). Sebagai contoh, apabila suatu tes bahasa Arab diteskan kepada individu-individu atau kelompok yang sama berulang kali (misalnya dua kali) dalam situasi yang sama kemudian skor yang diperoleh juga

ISSN: 2540-9417

relatif sama atau dalam bahasa statistiknya tingkat keajegannya itu signifikan, maka tes tersebut memiliki daya reliabilitas. Untuk mengetahui tingkat signifikansi keajegan suatu tes digunakan teknik analisis korelasi. Artinya, hasil tes pertama pada subjek yang sama dikorelasikan dengan hasil tes yang kedua. Apabila hasil dari kedua tes tersebut menunjukkan taraf signifikansi, maka dapat dikakatan, bahwa tes tersebut memiliki daya reliabel.

Karakteristik tes yang ketiga adalah tingkat kesulitan (*darajatush shu'ubah* atau *mustawa shu'ubatil ikhtibar*). Karakteristik ketiga ini mengacu pada pengertian bahwa tes itu tingkat kesulitannya tidak terlalu ekstrim. Ekstrimitas tingkat kesulitan ini berada pada apa yang disebut dengan "terlalu mudah" dan terlalu sulit". Tes dikatakan terlalu mudah manakalah setiap butir soal dalam suatu tes dijawab benar oleh hampir semua teste, sebaliknya, tes dikatakan terlalu sulit, manakala setiap butir soal dalam suatu tes dijawab salah oleh hampir semua teste. Menurut Djiwandono (1996), tingkat kesulitan yang dalam penghitungan sering diberi tanda ***p***, dapat dinyatakan dengan persentase (%). Cara penghitungan angka tingkat kesulitan dapat diperoleh melalui penghitungan sederhana, yaitu dengan rumus :

P = (JJB:JPT) x 100%.
P = Tingkat kesulitan butir tes
JJB = Jumlah jawaban benar
JPT = Jumlah peserta tes.

Berkaitan dengan hal ini, Oller (1979) menyatakan bahwa suatu butir tes dinyatakan layak jika indeks tingkat kesulitannya berkisar antara 0,15 sampai dengan 0,85. Ada pendapat lain yang mengatakan bahwa suatu butir tes dianggap baik bila mempunyai *p* antara 10% sampai 90% (Joni, 1986). Hal ini berarti, bahwa apabila indeks suatu butir tes di bawah 0,15 atau 0,10, maka butir tes tersebut tergolong terlalu sulit. Sebaliknya, apabila indeks suatu butir tes lebih dari 0,85 atau 0,90, maka butir tes tersebut tergolong terlalu mudah sehingga butir tes tersebut perlu direvisi atau tidak digunakan. Sebagai contoh, apabila suatu butir tes dijawab benar oleh tujuh peserta tes dari jumlah keseluruhan peserta sebanyak 30, maka tingkat kesulitan butir tes terebut adalah 23%, atau 0,23. Apabila suatu butir tes dijawab benar oleh dua peserta tes dari jumlah keseluruhan peserta sebanyak 30, maka tingkat kesulitan butir tes tersebut adalah 6,67% atau 0,067. Berdasarkan indeks tersebut (0,23), maka butir tes dengan angka tergolong baik. Sementara itu, tingkat kesukaran sebesar 6,67% menunjukkan bahwa butir tes tersebut tergolong sulit. Implikasinya, butir tes ini harus direvisi atau tidak digunakan.

Karakteristik yang keempat adalah daya beda (*darajatut tamyiz*). Sebuah tes dikatakan memiliki daya beda yang baik manakalah tes tersebut dapat membedakan antara siswa kelompok tinggi dan kelompok rendah. Menurut Joni (1986), suatu soal dikatakan mempunyai kemampuan diskriminasi (daya beda) yang benar apabila soal tersebut dijawab benar oleh lebih banyak anggota kelompok pintar (*upper group*) bila dibandingkan dengan anggota kelompok tidak pintar (*lower group*). Sebagai gambaran, apabila satu kelas terdiri dari 30 teste, maka diambil 27% sebagai kelompok atas dan 27% sebagai kelompok bawah. Dengan demikian ada 8 teste (dibulatkan) berdasarkan urutan skor sebagai kelompok atas (*upper group*), yaitu teste peringkat 1 sampai 8 dan 8 teste dari bawah berdasarkan urutan skor sebagai kelompok bawah (*lower group*), yaitu teste peringkat 23 sampai 30. Semakin tinggi

ISSN: 2540-9417

tingkat daya beda atau diskriminasi suatu butir tes, semakin tinggi pula kemampuannya untuk membedakan peserta yang pandai dari yang kurang pandai (Djiwandono, 1996). Kemampuan diksriminasi atau daya beda dapat dinyatakan dengan % dengan mempergunakan simbul *D* (Joni, 1986).

Untuk menentukan daya beda suatu butir tes dapat digunakan rumus berikut.

$$D = \frac{\sum \text{benar upper} - \sum \text{benar lower}}{\sum \text{kelompok (upper atau lower)}} \times 100\%$$

Dalam menentukan rentangan indeks daya beda dari suatu tes, dapat digunakan rumus yang dikemukakan oleh Djiwandono (1996:144) sebagai berikut.

0,50 atau lebih	: baik
antara 0,20 dan 0,50	: sedang
di bawah 0,20	: kurang
0	: tidak ada deskriminasi
- (negatif)	: negatif

Sebagai contoh, apabila suatu butir soal dijawab benar oleh enam teste dari kelompok atas dan seorang teste dari kelompok bawah, maka cara penghitungannya adalah sebagai berikut.

$$D = \frac{6-1}{8} \times 100 = 0{,}63$$

Dengan demikian, butir soal tersebut memiliki daya beda yang baik, karena indeks daya beda menunjukkan angka 0,63 (> 0,50). Ini berarti, bahwa butir tes ini dilihat dari *D*-nya dapat membedakan kelompok atas dan bawah.

C. PENILAIAN BAHASA ARAB DALAM KURIKULUM 2013

Rekonstruksi sebuah kurikulum sedikit banyak berdampak pada rekonstruksi bagian-bagiannya (komponen-komponennya) baik secara parsial maupun maupun secara kesekluruhan, termasuk rekonstruksi pada sistem penilaian. Secara substantif-akademik, sistem penilaian dalam kurikulum 2013 (K 13) dan Kurikulum sebelumnya (KBK dan KTSP) tidak jauh berbeda. Perbedaannya terletak pada porsi ranah atau domain yang dinilai. Apabila pada kurikulum sebelumnya ranah afektif (sikap) belum mendapatkan porsi yang eksplisit-aplikasional, maka penilaian dalam K13 ini, penilaian sikap/afektif, termasuk penilaian dalam pembelajaran bahasa Arab telah menempatkan penilaian sikap secara seimbang. Hal ini ditegaskan dalam K 13, bahwa Kompetensi Inti (KI)harus menggambarkan kualitas yang seimbang antara pencapaian *hard skills* dan *soft skills*. Bahkan dalam rumusan KI—pengganti istilah Standar Kompetensi (SK)—dan Kompetensi Dasar (KD), dan Indikator, butir-butir atau rumusan sikap (karakter) dikemukaan secara ekplisit dan ditempatkan pada urutan pertama.

KI dirancang dalam empat kelompok yang saling terkait yaitu berkenaan dengan sikap keagamaan (kompetensi inti 1), sikap sosial-edukasional (kompetensi 2), pengetahuan (kompetensi inti 3), dan penerapan pengetahuan (kompetensi 4). Keempat kelompok itu menjadi acuan dari Kompetensi Dasar dan harus dikembangkan dalam setiap peristiwa pembelajaran secara integratif. Kompetensi yang berkenaan dengan sikap keagamaan dan sosial-edukasional dikembangkan

ISSN: 2540-9417

secara tidak langsung (*indirect teaching*) yaitu pada waktu peserta didik belajar tentang pengetahuan (kompetensi kelompok 3) dan penerapan pengetahuan (kompetensi Inti kelompok 4).

Posisi penilaian dalam K 13 sebagaimana di atas mengimplikasikan penerapan penilaian dalam pembelajaran bahasa Arab yang multi dimensi. Dimensi ranah, dimensi proses, dan dimensi bentuk (instrumen). Penilaian dalan dimensi ranah mengacu pada ketiga taksonomi secara seimbang, yakni sikap, pengetahuan, dan keterampilan (kalau bukan psikomotor). Dimensi proses mengacu pada berbagai strategi penyelenggaraan penilaian dalam pembelajaran bahasa Arab. Dimensi bentuk (instrument) mengacu pada variasi instrumen yang digunakan.

Penilaian pada ranah sikap menggambarkan aspek integritas peserta didik dalam pembelajaran bahasa Arab. Integritas tersebut diindikasikan oleh prilaku mereka yang dalam KI diformulasikan dalam bentuk menghayati dan mengamalkan ajaran agama, mengembangkan perilaku jujur, disiplin, tanggungjawab, peduli, santun, kerjasama, responsif dan proaktif, dan perilaku sejenisnya. Pada ranah pengetahuan, kompetensi yang dinilai menggambarkan pemahaman mereka terhadap sistem unsur-unsur atau komponen bahasa, sedangkan kompetensi pada tataran keterampilan menggambarkan kemampuan mereka dalam berbahasa Arab (*istima'*, *kalam*. *Qira'ah*, dan *kitabah*).

Dari dimensi strategi atau cara, penilaian dalam K 13 mengimplikasikan kebervariasian cara menilai kompetensi berbahasa Arab peserta didik. Kebervariasian cara dimaksudkan bahwa pendidik menilai kompetensi berbahasa Arab mereka tidak hanya terbatas pada UTS, UAS, atau setelah KD tertentu yang penyelenggaraannya dengan tes tertulis. Akan tetapi, penilaian dapat dilakukan pada saat proses pembelajaran yang penyelenggaraannya dalam bentuk tes lisan dan tulis.

Dari dimensi bentuk atau instrumen, penilaian dilakukan dengan menggunakan berbagai instrumen. Keberagaman instrumen penilaian dalam pembelajaran bahasa Arab ini dimaksudkan agar kompetensi berbahasa Arab yang dinilai menggambarkan kompetensi menyeluruh dan menggambarkan pula kompetensi yang sebenarnya atau senyatanya, bukan kompetensi artifisial. Bentuk-bentuk instrumen yang dapat digunakan—selain tes (UTS,UAS) misalnya portofolio, jurnal, lembar observasi, tes performansi,dan lain lain. Model penilaian inilah yang disebut dengan penilaian otentikatau *authentic assessment*atau *attaqwim al-waqi'i* (O'Malley, 1996).

D. MENUJU TES YANG SAHIH

Sebagaimana dikemukakan, bahwa kesahihan merupakan syarat utama dalam penyusunan tes yang baik. Bahkan kesahihan ini lebih prinsip dibandingkan dengan karakteristik lainnya. Oleh karena itu, tidaklah berlebihan apabila dikatakan bahwa tes yang sahih itu reliabel. Untuk mewujudkan sebuah tes yang sahih diperlukan langkah-langkah tertentu agar tes yang disusun memiliki daya validitas tinggi baik dari aspek validitas isi (*content validity* atau *shidqul muhtawa*) maupun validitas tampak luar (*face validity* atau *ash-shidqudz dzahiry*). Langkah-langkah yanag memungkinkan terwujudunya tes bahasa Arab yang baik, khususnya tes untuk mengukur hasi belajar adalah sebagai berikut:

ISSN: 2540-9417

Pertama menetapkan tujuan penilaian. Tes yang disusun baik yang terkait dengan cara, bentuk tes, dan isinya harus mengacu pada tujuan penyelenggaraan tes itu sendiri. Tujuan tes ini dapat mengacu pada kurikulum bahasa Arab atau mengacu pada tujuan yang didesain oleh guru atau lembaga itu sendiri. Dalam kurikulum, khususnya K 13, tujuan pengajaran bahasa Arab tercermin dalam rumusan pada kompetensi inti dan kompetensi dasar. Apabila tujuan tes bahasa Arab, misalnya di MI kelas III itu untuk mengukur kemampuan kosa kata, maka butir-butir soal yang disusun tentunya mengukur kemampuan kosa kata mulai dari makna kata sampai penggunaan kosa kata secara praktis.

Kedua menetapkan materi tes. Yang dimaksudkan menetapkan materi tes di sini adalah menetapkan materi tes bahasa Arab yang meliputi komponen bahasa dan atau *maharah* yang akan diukur sesuai dengan tujuan yang ditetapkan. Misalnya materi kosa kata, pola kalimat, *maharah istima'*, *maharah kalam*, *maharah qira'ah* dan *kitabah*. Tentunya materi-materi ini tidak serta merta secara keseluruhan dan simultan diteskan sekaligus. Akan tetapi, kemungkinan dipilih *maharah* atau komponen tertentu sesuai dengan tujuan pembelajaran bahasa Arab di madrasah/sekolah. Apabila tujuan tesnya itu untuk mengukur kemampuan kosa kata, membaca, dan menulis terbimbing, maka materi tesnya juga meliputi ketiga kemampuan tersebut. Akan kurang sahih, manakala dimasukkan tes *maharah hiwar*, apalagi pelaksanannya dalam bentuk tulis (bukan aktifitas lisan).

Yang sering "terabaikan" oleh guru adalah tes *maharah kalam*. Tes *maharah kalam* ini, baik dalam cara penyelenggaraannya dan instrumennya tentu berbeda dengan penyelenggaraan tes tulis. Tes *maharah kalam* ini diselenggarakan dalam aktifitas komunkasi bahasa Arab secara lisan baik dalam bentuk dialog maupun monolog. Sistem penyelenggarannya juga tidak bersifat klasikal, melainka lebih bersifat individual atau kelompok-individual dan membutuhkan waktu yang relatif lama. Oleh karena itu, guru "agak menghindar" penyenggaraan tes *maharah kalam*.

Ketiga memilih *dars* atau *maudlu'* yang representatif. Pemilihan dan pemilahan *dars* atau *maudlu'* yang representatif ini dilakukan agar materi yang dipilih sebagai bahan tes tidak bias. Untuk menghindari kebiasan dalam pemilihan dan pemilahan *dars* ini, guru tidak harus menjadikan semua materi atau *dars* yang dipelajari itu dijadikan sebagai bahan tes, apalagi materi atau *dars* yang dipelajari banyak, sementara waktu yang teresdia untuk menjawab relatif terbatas. Dalam konteks inilah, guru bahasa Arab mempunyai otoritas untuk menjustifikasi pemilihan materi tes secara "*sampling*" yang merepresentasikan materi secara kesulurhan.

Keempat menentukan bentuk soal dan cara penyelenggaraannya. Sebagaimana diketahui bersama, bahwa bentuk tes itu bermacam-macam, termasuk instrumen yang digunakan. Cara penyelenggaraannya juga bisa bervariasi, misalnya tes lisan atau tulis sesuai dengan tujuan tes itu sendiri. Ada kecendrungan bahwa bentuk tes bahasa Arab di sekolah dan atau di madrasah diseragamkan dengan tes matapelajaran yang lain, termasuk jumlah butir soalnya. Misalnya bentuk tesnya itu objektif dengan variasi pilihan ganda (30 butir soal) dan salah benar (10 soal), serta bentuk tes esai (5 butir soal).

Kelima menentukan jumlah butir soal. Perihal yang harus dipertimbangkan dalam menentukan jumlah butir tes adalah alokasi waktu yang tersedia untuk penyelenggaraan tes. Untuk menentukan berapa jumlah butir tes yang harus

ISSN: 2540-9417

disusun sesuai dengan waktu yang tersedia memang tidak ada batasan yang pasti. Akan tetapi, guru dengan otoritas kepengalamannya yang memahami kondisi objektif siswanya dan substansi kurikulumakan dapat menentukan jumlah butir tes yang sesuai dengan waktu yang tersedia. Masing-masing komponen maupun keterampilan berbahasa Arab mendapatkan porsi jumlah butir soal yang proporsional sesuai dengan bobot tujuan dan perioritas pembelajaran bahasa Arab di sekolah atau madrasah. Apabila dalam kurikulum terdapat penjelasan alokasi waktu persemester atau pertahun untuk masing-masin komponen dan atau masing-masing *maharah*, maka alokasi waktu itulah yang dpat dijadikan acuan untuk menentukan jumlah butir soal. Untuk penghitungan jumlah buir soal dapat digunakan rumus sebagai berikut.

Rumus I $$= \frac{\sum \text{jam per aspek}}{\sum \text{jam keseluruhan}} \text{x} 100$$

Selanjutnya hasil pada rumus 1 dikembangkan pada rumus 2 yaitu (Hasil penghitungan rumus I:100) x jumlah butir soal. Misalnya penghitungan presentase butir soal untuk aspek *maharah qiraah* 37% dengan jumlah butir soal 40 (pilihan ganda), maka jumlah soal untuk *maharah qiraah* adalah(37:100) x 40=14,8 dibulatkan menjdi 15 butir soal.

Cara lain yang paling sederhana misalnya guru memberikan presentasi per aspek komponen bahasa dan *maharah* berdasarkan perioritas tujuan pembelajaran bahasa Arab di madrasah atau di sekolah. Misalnya butir tes *maharah qiraah* 40%, dan yang 60% untuk *maharah* lainnya. Demikian pula, apabila bentuk soal yang dikembangkan dalam tes bervariasi, maka guru juga harus mempertimbangkan proporsi untuk masing-masin komponen dan *maharah*. Intinya, jumlah butir soal yang dikembangkan dalam tes tidak bias dan tidak ada unsur subjektifitas guru tanpa memperhatikan esensi tujuan pembelajaran bahasa Arab dan penekanan kompetensi yang dikuasai oleh peserta didik.

Keenam membuat kisi-kisi tes. Kisi-kisi tes merupakan gambaran umum tentang materi dan kompetensi yang dijadikan acuan dalam penyusunan tes. Setiap butir soal yang dibuat harus mengacu pada kisi-kisi tes. Kesesuaian butir-butir soal dengan kisi-kisi yang ditetapkan mengindikasikan kesahihan sebuah tes (kesahihan logis). Berdasarkan pengalaman dalam menelaah tugas-tugas mahasiswa (S2 PBA) seringkali diketemukan ketidaksesuaian antara butir soal dengan kisi-kisi yang dijadikan acuan. Misalnya, dalam kisi-kisi disebutkan, bahwa butir soal no. 6 mengukur kemampuan menemukan ide pokok dalam sebuah paragraf, tetapi butir soalnya mengukur menemukan judul bacaan.

Tabel 1 berikut ini merupakan contoh kisi-kisi tes bahasa Arab, misalnyadi MTs (selain *maharah kalam*) dalam bentuk soal Pilihan Ganda (PG) sebanyak 40 butir soal dan esai sebanyak 5 butir soal *maharah qiraah* dan 5 butir soal *maharah kitabah*dengan proporsi jumlah butir soal PG sebagai berikut: penggunaan kosa kata 20%= 8 butir soal), penggunaan *tarkib* (20%= 8 butir soal), pemahaman teks (40% = 16), dan menulis terbimbing (20% = 8).

ISSN: 2540-9417

Tabel 1: Contoh Kisi-Kisi Tes Bahasa Arab Non-*Istima* dan *Kalam*

Kompetensi Dasar	Indikator	Bentuk soal Pilihan Ganda		Bentuk Soal Esai	
		No. Soal	Jumlah	No. Soal	Jumlah
1. Penggunaan Kosa Kata	Arti kata	1,2,	2	-	-
	Persamaan kata	3,4	2	-	-
	Penggunaan kata	5,6,7,8,	4	-	-
2. Penggunaan *tarkib*	*istiham*	9, 10	2	-	-
	dhomir	11, 12	2	-	-
	Fi'l – fa'il	13, 14	2	-	-
	Mubtada-khabar	15, 16	2	-	-
3. Memahami teks	Menemukan informasi tersurat	17, 18, 19, 25, 26, 27	6	1,2,3	3
	Menemukan informasi tersirat	20, 21, 23, 24,	4	4, 5	2
	Menemukan topik	22, 29	2	-	-
	Menerjemahkan	28, 30, 31, 32	4	-	-
4. Menulis terbimbing	Menyusun kata menjadi kalimat	33,34,35	3	1, 2	2
	Membuat kalimat tanya dari kalimat informatif	36, 37,	2	3	1
	Mendeskripsikan gambar tunggal	38, 39,40	3	4, 5	2
Jumlah			40		

Bagaimana halnya dengan kisi-kisi untuk *maharah istima'* dan *maharah kalam*. Untuk kisi-kisi *maharah istima*, rumusan kompetensi dan indikatornya tidak jauh berbeda dengan *maharah qira'ah*, perbedaannya terletak pada pemahaman teks (*fahmul maqru'*) dan memahami wacana lisan (*fahmul masmu'*). Sementara itu, kisi-kisi untuk *maharah kalam* disusun atas dasar fungsi komunikatif sesuai dengan topik komunikasi dan bersifat *direct- observation*. Istilah lain yang dapat digunakan adalah tes performansi yang implementasinya bisa berbentuk aktifitas dialog atau monolog. Dialog itu sendiri bisa antarsiswa dan bisa antara guru dengan siswa (wawancara). Tabl 2 berikut ini merupakan contoh kisi-kisi tes performansi (*maharah kalam*)

Tabel 2: Conto Kisi-Kisi Tes Bahasa Arab *Maharah Kalam*

Kompetensi Dasar	**Indikator**	**Kegiatan**	**Bentuk**
Mampu melakukan	Saling	dialog	Bermain peran

ISSN: 2540-9417

Perkenalan	memperkenalkan diri sendiri (nama, alamat, no. Hp, tempat dan tanggal lahir, hobi, profesi, dst).		
	Memperkenalkan orang lain (nama, alamat, tempat dan tanggal lahir, hobi, profesi, dst).	monolog	bercerita
	Memperkenalkan anggota keluarga (nama anggota keluarga, posisi dalam keluarga, profesi anggota keluarga, dst).	monolog	bercerita
Mampu menceritakan kegiatan sehari hari	Saling menceritakan kegiatan sehari hari di rumah	dialog	Bermain peran
	Menceritakan kegiatan sehari hari di sekolah	monolog	Wawancara dengan guru

Sementara itu, komponen atau aspek yang dinilai dalam tes performansi (*maharah kalam*)dapat mengadopsi model penilaian yang dikutip oleh Asrori, dkk. (2006) sebagaimana pada tabel 3 berikut.

Tabel 3: Format Standar Penilaian Maharah Kalam

Nama Siswa	: ……..	**Tanggal**	: ……
Korektor	: ……..	**Skor/Nilai**	: ……

Nilai	**Karakter Ujaran**:
5	Sedikit sekali kesalahan ujar (*speech defect*) yang muncul
4	Kesalahan ujaran lebih terlihat, tetapi masih dapat dimengerti dengan jelas
3	Terdapat kesalahan ujar yang mengundang perhatian lebih dalam menyimak, bahkan terkadang kurang dapat dipahami.
2	Sulit dipahami karena kesalahan ujar, sehingga harus banyak mengulang apa yang dikatakannya
1	Kesalahan ujar yang muncul mengakibatkan perkataannya tidak bisa dipahami sama sekali
Nilai	**Qowa'id**:
5	Kesalahan gramatikal dan susunan kalimat sedikit sekali, bahkan tidak terlihat
4	Terjadi kesalahan gramatikal, namun tidak menyebabkan kaburnya arti
3	Kadang kala terjadi kesalahan dan menyebabkan kaburnya arti
2	Kesalahan gramatikal menyebabkan sulit dipahami, dan siswa hanya mengulang-ulang satu bentuk atau pola kalimat
1	Banyak terjadi kesalahan arti karena kesalahan gramatikal yang fatal sampai perkataannya tidak dapat dipahami sama sekali

ISSN: 2540-9417

Nilai	Kosakata:
5	Menggunakan kosakata atau idiom sebagaimana para penutur asli dengan sempurna
4	Kadang kala menggunakan idiom yang kurang tepat atau terpaksa mengulang ide-ide sebelumnya karena kosakata yang dimiliki tidak dapat membantunya
3	Berkali-kali menggunakan kata-kata yang salah. Perkataannya sedikit karena perbendaharaan kosakata yang dimiliki terbatas.
2	Kesalahan dan minimnya jumlah kosakata yang digunakan menyebabkan perkataannya sangat sulit dipahami
1	Kosakata yang dimiliki sangat terbatas sehingga tidak mampu berbicara sama sekali
Nilai	**Kelancaran:**
5	Kelancaran berbicara siswa sama dengan penutur asli
4	Tempo berbicara terlihat agak berkurang karena masalah-masalah kebahasaan
3	Antara tempo dan kelancaran makin terpengaruh oleh masalah-masalah kebahasaan
2	Selalu mengulang-ulang (gagap, ragu). Dirinya kadang terpaksa diam karena keterbatasan bahasa.
1	Pembicaraan tersendat-sendat/terputus-putus sampai tidak mungkin dilakukan dialog
Nilai	**Pemahaman:**
5	Terlihat memahami semuanya tanpa kesulitan
4	Dalam tempo normal, mampu menguasai semuanya. Namun terkadang siswa minta mengulang perkataan yang disampaikan padanya
3	Mengerti sebagian besar apa yang dikatakan jika tempo lebih lambat dari biasanya dengan beberapa pengulangan
2	Banyak mengalami kesulitan dalam mengikuti pembicaraan sehingga siswa hanya mampu memahami jika disampaikan dalam tempo lambat dengan banyak pengulangan
1	Tidak mampu memahami apa yang diucapakan padanya kendati dalam percakapan paling sederhana dan mudah.

Keterangan: rincian 1 = gagal, 2 = kurang, 3 = cukup, 4 = baik, dan 5 = memuaskan.

Sementara itu, Harris (1969), secara global menegaskan bahwa berbicara itu merupakan keterampilan yang sangat kompleks yang mempersyaratkan penggunaan berbagai kemampuan secara simultan. Kemampuan tersebut meliputi: (a) pelafalan (yang mencakup ciri-ciri segmental-vokal dan konsonan, serta pola tekanan dan intonasi), (b) tatabahasa, (c) kosa kata, (d) kelancaran (*fluency*), dan (e) pemahaman (kemampuan merespon terhadap suatu ujaran secara baik).

Kreteria lain yang dapat dijadikan acuan dalam penilaian *maharah kalam* misalnya pada tabel 4 berikut ini.

Tabel 4: Aspek yang dinilai dalam *Maharah Kalam*

No.	Aspek yang dinilai	Skor Maksimal
1	Kelancaran	20
2	Ketepatan ujaran	15
3	Struktur	15
4	Pilihan Kosa Kata	15
5	Kelengkapan isi	25
6	Penampilan	10
	Jumlah	100

ISSN: 2540-9417

Ketujuh menyusun butir tes berdasarkan kisi-kisi. Dalam penyusunan butir soal ini, ada rambu-rambu yang sebaiknya diperhatikan oleh guru atau pembuat tes, yaitubahasa yang digunakan jelas dan lugas, tidak ambigu, substansi pertanyaan fokus pada permasalahan tertentu. Khusus untuk soal yang berbentuk pilihan ganda harus diperhatikan hal-hal sebagai berikut. (a) ***stem*** **(**pernyataan pokok**)** pada setiap butir tes (terutama butir tes pilihan ganda atau salah-benar) hanya berisi satu permasalahan, (b) panjang kalimat untuk setiap option (khusus untuk butir soal pilihan ganda) relatif sama. Hal ini dimaksudkan untuk menghindari adanya kemungkinan teste memilih ***option*** yang paling panjang sebagai jawaban yang paling benar, (c) letak jawaban yang benar (khusus untuk butir tes pilihan ganda maupun salah-benar) disusun secara acak (tidak linier). Artinya, harus dihindari letak jawaban benar yang berpola, misalnya (dalam soal pilihan ganda) berpola ab, ac, dan ad atau berpola aa, bb, cc, dan dd. (Asrori, dkk. 2012).

Kedelapan mereview tes yang sudah disusun. Tes yang sudah disusun dengan memperhatikan kisi-kisi yang ada sebaiknya ditelaah kembali baik secara individual maupun melalui tim pereview (bila ada). Melalui telaah ulang ini, akan dihasilkan sebuah tes yang sahih secara substansial dan sahih secara struktural. Sebuah tes yang kurang memenuhi kreteria kesahihan baik secara substansial maupun struktural akan berdampak pada kebingunngan peserta didik dalam menjawab soal. Apabila demikian halnya, maka tes ini masuk pada katagori tes yang lemah (*dho'if*).

SIMPULAN

Berdasarkan paparan di atas, simpulan yang dapat dikemukakan adalah sebagai berikut: (1) Salah satu tugas utama guru sebagai pengambil keputusan dalam pembelajaran adalah penilaian. (2) Secara fungsional, penilaian tidak hanya berfungsi untuk mengetahui keberhasilan dan proses pembelajaran, tetapi berfungsi sebagai *feed back* untuk perbaikan pembelajaran dan salah satu kegiatan penilaian adalah Tes. (3) Salah satu karakteristik tes yang baik adalah sahih. Kesahihan atau validitas sebuah tes terkait dengan apakah butir-butir tes yang disusun mengukur yang seharusnya diukur. (4) untuk mewujudkan tes yang baik diperlukan langkah-langkah: menetapkan tujuan penilaian, menetapkan materi tes, memilih *dars* atau *maudlu'* yang representatif, menentukan bentuk soal dan cara penyelenggaraannya, menentukan jumlah butir soal, membuat kisi-kisi, menyusun butir tes berdasarkan kisi-kisi, dan mereview tes yang sudah disusun.

DAFTAR RUJUKAN

Al-khuli, Muhammad Ali. 2000. *Al-ikhtbarat Al-Lughawiyyah*. Shuwailih (Al-Urdun): Darul Falah.

Asrori, Imam., Ainin, Moh., dan Tohir, Moh. 2012. *Evaluasi dalam Pembelajaran Bahasa Arab*. Malang: Misykat.

Cooper, James, M. 1977. The Teacher as Decision Maker Dalam james Cooper (Ed.), *Classroom Teaching skills: A Handbook*. Masschusetts. D.C. Heath and Company.

Djiwandono, M. Soenardi. 1996. *Tes Bahasa dalam Pengajaran*. Bandung: ITB.

Duwaidari, Roja' Wahid. 2000. *Al-bahtsul Ilmi: Asaasiyyatuhu An-Nadzariyyah wa mumarasatul amaliyyah*. Lubnan: Darul Fikri.

ISSN: 2540-9417

Gunning, Thomas G. 1998. *Assessing and Correction* Reading and Writing Difficulties. Boston: Allyn and Bacon.

Harris, David P. 1969. *Testing English as a Second Language*. New York: McGraw-Hill Book Company.

Joni, T. Raka. 1986. *Pengukuran dan Penilaian Pendidikan*. Surabaya: Karya Anda.

McNamara, Tim. 2008. *Language Testing*. (H.G. Widdowson, Ed.). New York: Oxfor University Press.

Oller, Kohn W. 1979. *Language Tests at School*. London: Longman Group Ltd.

O'Malley, J.M. dan Pierce, L.V. 1996. *Authentic Assessment for Englisgh Language Learners*. Wesley: Addison Wesley Publishing Company.