

ANALISIS BUTIR TES TOAFL
(*TEST OF ARABIC AS A FOREIGN LANGUAGE*)
UIN MALIKI MALANG

Oleh: Syaiful Mustofa
UIN Maulana Malik Ibrahim Malang
Email: saifulmustofa@pba.uin-malang.ac.id

A. Latar Belakang

Berdasarkan rekomendasi hasil rapat kerja UIN Malang tentang sosialisasi program dan anggaran tahun 2016 dan perancangan kegiatan tahun 2017 di Prigen Pasuruan pada tanggal 19-21 pebruari 2016 lalu, menyebutkan bahwa tes TOAFL akan diberlakukan untuk calon wisudawan mahasiswa S1, S2 dan S3 di lingkungan UIN Malang. Tujuannya adalah mengukur kebahasaan yang telah diperoleh peserta didik setelah menempuh atau memperoleh pengalaman belajar dalam waktu tertentu, disamping juga sebagai pendamping ijazah.

TOAFL (*Test of Arabic as a Foreign Language*) bertujuan untuk meningkatkan standar mutu kelulusan secara akurat dan jelas, sehingga tingkat kemampuan bahasa Arab lulusan UIN, IAIN, STAIN dan PTAIS dapat diukur dengan standar tertentu secara pasti.

Penggunaan TOAFL sebagai materi tes didasarkan pada kebutuhan akademis bahwa mahasiswa dituntut mampu dan memiliki standar tertentu dalam berbahasa Arab, akan tetapi tes TOAFL versi Arab ini belum banyak dikembangkan di dunia Arab dan di Indonesia. Padahal dalam perkembangannya TOAFL sekarang ini digunakan sebagai persyaratan materi tes masuk di berbagai program pascasarjana perguruan tinggi Indonesia maupun di luar negeri, program Kemenag RI 5000 doktor, petugas haji, dll.

Tes TOAFL dilaksanakan di beberapa kampus dengan butir soal tesnya telah terdaftar di HAKI (hak kekayaan intelektual) dan mempunyai lisensi tertentu termasuk yang akan dilakukan oleh pusat pengembangan bahasa (P2B) UIN Malang, sebelum tes TOAFL didaftarkan ke HAKI maka perlu ada evaluasi dan analisis butir tes TOAFL sehingga penelitian ini sangat signifikan.

Tes TOAFL salah satu kebutuhan inti dalam kehidupan akademis. Tetapi, mahasiswa perguruan tinggi di lingkungan PTAIN dan PTAIS masih banyak yang belum diukur standar mutu kelulusannya secara akurat dan jelas setelah menempuh atau memperoleh pengalaman belajar di kampusnya, hal ini disebabkan tidak semua

perguruan tinggi menyediakan tes TOAFL. Pusat pengembangan bahasa (P2B) UIN Malang telah mengembangkan tes TOAFL sejak tahun 2009 lalu, akan tetapi butir tes belum di *upgrade* dan belum didaftarkan ke HAKI. Peneliti perlu melakukan evaluasi dan analisis butir tes TOAFL agar sesuai dengan kebutuhan pembelajar bahasa Arab saat ini. Berdasarkan pokok masalah ini, maka penelitian dengan topik di atas sangat *urgen* dilakukan.

Obyek penelitian ini adalah tes TOAFL yang dikembangkan Pusat Pengembangan Bahasa (P2B) UIN Maliki Malang tahun 2016. Hal ini didasarkan atas pertimbangan bahwa tes TOAFL UIN Maliki Malang menjadi rujukan mahasiswa seluruh Indonesia untuk mendapatkan sertifikat. Butir tes TOAFL UIN Maliki Malang sudah di *upgrade* pada bulan januari dan pebruari 2016 lalu, akan tetapi butir tes tersebut belum diukur validitas, realibilitas, tingkat daya beda, dan tingkat kesukaran sehingga belum diketahui standart validitas, realibilitas, tingkat daya beda dan tingkat kesukaran butir tes TOAFL tersebut.

B. Kajian Pustaka

1. Konsep Analisis Butir Tes

Analisis butir tes merupakan kegiatan penting dalam upaya memperoleh instrumen yang berkategori baik. Analisis ini meliputi (1) menentukan validitas dan realibilitas tes, dan (2) analisis butir (*item analysis*).

Analisis terhadap butir tes yang telah di jawab siswa suatu kelas mempunyai dua tujuan yakni (1) jawaban-jawaban soal tersebut merupakan informasi diagnostik untuk meneliti pelajaran dari awal dan kegagalan belajarnya, serta selanjutnya untuk membimbing ke arah cara belajar yang lebih baik; dan (1) jawaban terhadap soal dan perbaikan (*review*) soal-soal yang didasarkan atas jawaban tersebut merupakan dasar bagi penyiapan tes yang lebih baik.¹

Dengan melakukan analisis butir sedikitnya kita dapat mengetahui empat hal penting, yakni:

- bagaimana taraf kekuasaan setiap butir tes.
- apakah setiap soal memiliki validitas dan realibitas tinggi.
- apakah setiap soal memiliki daya pembeda baik.
- sejauh mana tiap butir tes dapat mengukur hasil pembelajaran.

¹Thorndike dan Hagen (Purwanto, 1992)

2. Tes sebagai Instrumen Evaluasi

Tes merupakan instrumen atau prosedur yang sistematis untuk mengukur kemampuan tertentu, dirancang dan dilaksanakan kepada pembelajar pada waktu dan tempat tertentu serta dalam kondisi yang memenuhi syarat-syarat yang jelas.²

Secara terminology bahasa Arab, tes disebut *ikhtibâr*. Salah satu instrumen evaluasi pembelajaran bahasa Arab adalah tes TOAFL. Tes ini dirancang dan disusun sesuai dengan tujuan dan sasaran pembelajar itu sendiri untuk mengukur mutu kebahasaan mahasiswa secara akurat dan jelas dengan standar secara pasti. Mengukur tingkat keberhasilan pembelajar merupakan salah satu tugas utama pengajar karena memang berkaitan dengan tugas edukatifnya, yakni memberi evaluasi dan penilaian terhadap pemerolehan hasil belajar peserta didik.³

Tes kebahasaan merupakan sejumlah prosedur dan instrument yang didesain secara sistematis, digunakan oleh seseorang atau tenaga pengajar dalam mengamati dan mengetahui performa dan kompetensi salah satu keterampilan bahasa peserta didik atau secara keseluruhannya, sesuai dengan ukuran tertentu untuk mencapai maksud dan tujuan tertentu pula. Pengerjaan tes tergantung petunjuk yang diberikan, misalnya: melingkari, memberi tanda silang pada jawaban yang paling benar, mencoret jawaban yang salah, menjelaskan, mengisi titik titik, dan sebagainya.

3. Bentuk Tes Bahasa Arab

Tes kebahasaan itu sangat beragam, tergantung pada tujuan, kepentingan, cara pemeriksaan dan ruang lingkungannya. Dari segi tujuannya, tes kebahasaan dapat diklasifikasi menjadi tiga, yaitu: 1) tes pemerolehan atau tes prestasi (*achievement test, al-ikhtibar al-tahshili*), 2) tes profisiensi (*proficiency test, ikhtibar al-ijadah aw al kafaah*), 3) tes kesiapan berbahasa (*language aptitude test, ikhtibar al-isti'dad al lughawi*) atau tes prediksi (*predictive test, al ikhtibar al tanabbu'*).⁴

Tes pemerolehan bahasa adalah tes yang dimaksudkan menguji apa yang telah diperoleh peserta didik setelah menempuh atau memperoleh pengalaman pendidikan dalam waktu tertentu. Tes ini terkait dengan kurikulum dan bahan ajar yang digunakan oleh lembaga pendidikan, biasanya tes dilakukan dalam bentuk ujian pada pertengahan

² Suharso, & Ana Retnoningsih, *Kamus Besar Bahasa Indonesia*, Semarang: CV Widya Karya, 2005, cet-1
³ ماهر إسماعيل صبر محمد يوسف، ٢٠٠١٣. *التقويم التربوي: أسسه وإجراءاته*. الرياض: مكتبة الرشد.
⁴ رشد أحمد طعيمة، ٢٠٠١. *مناهج تدريس اللغة العربية بالتعليم الأساسي*. القاهرة: دار الفكر العربي
محمد عبد الخالق، ١٩٨٩. *اختبارات اللغة*. الرياض: جامعة الملك سعود. الطباعة الأولى

atau akhir semester.

Tes profisiensi adalah tes yang tidak dimaksudkan untuk menguji pemerolehan bahasa peserta didik, dan tidak terkait dengan kurikulum, buku ajar dan masa program belajar tertentu, melainkan menguji kemampuan dan keterampilan bahasa peserta didik secara umum. Yang termasuk jenis tes ini adalah TOEFL (*Test of English as a Foreign Language*) atau TOAFL (*Test of Arabic as a Foreign Language*).

Tes kesiapan atau tes prediksi adalah tes yang dimaksudkan untuk menentukan tingkat kesiapan peserta didik untuk belajar bahasa kedua, dan memprediksi kemajuan yang akan dicapai peserta didik. Tes ini juga mengukur aspek *audio-visual* peserta didik, terutama mengukur kemampuannya dalam membedakan berbagai *tarâkîb lugawiyyah*.

Dari segi pembuatnya, tes dapat dibagi menjadi dua, yaitu: tes standar (*al-ikhtibâr al-muwahhadah*) dan tes tenaga pengajar (*ikhtibâr al-mu'allim*). Yang pertama adalah tes yang dibuat oleh lembaga tertentu, dengan standar tertentu pula, untuk dipergunakan dalam skala yang luas, misalnya: tes bahasa Arab untuk seluruh kelas III Madrasah Aliyah dalam ujian akhir di wilayah kota Malang. Sedangkan yang kedua adalah tes yang dibuat oleh tenaga pengajar untuk diujikan kepada peserta didiknya sendiri, dan bertujuan untuk mengetahui tingkat penguasaan bahasa yang telah dipelajarinya.

Sementara itu, dari segi skoringnya, tes dapat dibagi menjadi dua, yaitu: tes essay atau tes subyektif dan tes obyektif. Yang pertama adalah tes yang dirancang sedemikian rupa, sehingga peserta didik memiliki kebebasan dalam memilih dan menentukan jawaban dalam bentuk uraian. Tes ini disebut subyektif karena jawaban peserta didik maupun koreksi yang diberikan oleh tenaga pengajar bersifat subyektif. Sedangkan yang kedua adalah tes yang butirnya dapat dijawab dengan memilih jawaban yang sudah tersedia, sehingga peserta didik menampilkan keseragaman data, baik yang menjawab benar maupun yang menjawab salah. Tes ini disebut obyektif karena pilihan jawaban bersifat pasti dan tertutup, tidak membuka peluang bagi peserta didik untuk memilih selain dari pilihan jawaban yang sudah ditentukan; demikian juga penilai juga tidak mungkin memberikan skoring yang menyimpang dari pilihan jawaban yang benar.

Setidaknya ada empat bentuk tes obyektif, yaitu: pilihan ganda (*al-ikhtiyâr min muta'addid, multiple choise*), pilihan benar-salah (*ikhtiyâr al-shawâb wa al-khatha'*),

mencari pasangan (*al-muzâwajah, matching*), dan melengkapi isian (*al-takmilah, completion*) dengan jawaban yang bersifat tertutup.

Dari segi cara dan bentuk pengujiannya, tes dapat dibagi menjadi dua: tes lisan (*ikhtibâr syafawî*) dan tes tulis (*ikhtibâr tahrîrî*). Yang pertama adalah tes yang soal dan jawabannya diberikan secara lisan, sebaliknya yang kedua adalah tes yang soal dan jawabannya diberikan dalam bentuk tulis. Tes lisan dapat digunakan, terutama untuk menguji keterampilan berbicara (*mahârat al-kalâm*), membaca dan ekspresi verbal (*ta'bir syafawî*). Sedangkan tes tulis dapat digunakan untuk menguji cabang-cabang kebahasaaraban yang kurang cocok diujikan secara lisan, seperti: materi nahwu, tarjamah tahrîriyyah (*tarjamah tulis*), *insyâ'*, dan sebagainya.⁵

Aplikasi tes, dalam berbagai bentuk dan jenisnya tersebut, dalam pembelajaran bahasa Arab dapat disesuaikan dengan karakteristik materi yang akan diujikan. Materi *istimâ'* berbeda dengan materi *qawâ'id* dan *insyâ'*. Demikian juga alat dan media yang digunakan. Tes keterampilan menyimak (*ikhtibâr al-istimâ'*), misalnya, idealnya dilakukan dalam laboratorium bahasa dengan menggunakan *tape recorde* dan *earphone*, atau sekurang-kurangnya didukung oleh rekaman kaset yang dibunyikan melalui tape, seperti halnya tes *listening* dalam TOEFL atau TOAFL.

Tes *mufradât* juga dikembangkan dengan penuh variasi, tidak hanya berupa mencari sinonim dan antonim kata, melainkan juga dapat berupa mendefinisikan sesuatu, menyebut profesi, mencari salah kata yang asing dari suatu kelompok kata, dan sebagainya.

Penyusunan tes harus sesuai dengan norma-norma berikut. Pertama, butir-butir atau kalimat soal hendaknya hanya disesuaikan dengan tujuan khusus yang telah ditetapkan. Misalnya saja, jika kalimat soal ditujukan untuk menguji arti *mufradât* dalam sebuah kalimat, maka alternatif jawaban –jika berbentuk pilihan ganda— hendaknya tidak bias dengan unsur nahwu atau sharaf. Kedua, soal yang dibuat hendaknya sesuai dengan tingkat kemampuan peserta didik, terutama jika berbentuk tes pemerolehan. Ketiga, penyusunan tes hendaknya disertai petunjuk yang jelas, baik mengenai cara dan tempat menjawabnya serta lamanya waktu yang disediakan. Keempat, redaksi atau rumusan masing-masing soal harus jelas, tidak bersayap dan multiinterpretasi, terukur, dan diskriminatif. Kelima, waktu yang diberikan untuk menjawab soal harus sebanding

⁵ رشيد أحمد طعيمة ومحمد سيد مناء. ٢٠٠٠. *تدريس العربية في التعليم العام: نظريات وتجارب*. القاهرة: دار الفكر العربي

dengan tingkat kesulitan dan banyak soal. Keenam, skoring penilaian harus obyektif berdasarkan proporsi yang ditetapkan, bukan berdasarkan rekaan, dan jauh dari subyektivitas penilai.⁶

4. TOAFL (Test of Arabic As a Foreign Language)

Uioouio

TOAFL adalah singkatan dari "Test of Arabic as Foreign Language". Nama ini diilhami oleh TOEFL (*Test of English as a Foreign Language*), yang telah ada lebih dahulu. Penamaan ini memang dimaksudkan agar TOAFL lebih mudah diucapkan dan dikenal oleh banyak orang, meskipun terkesan "menyerupai" TOEFL.

TOAFL dilatarbelakangi oleh upaya serius untuk meningkatkan standar mutu kelulusan secara akurat dan jelas, sehingga tingkat kemampuan bahasa Arab lulusan PTKIN dapat diukur dengan standar tertentu secara pasti. Penyusunan TOAFL juga disemangati oleh usaha "memasukkan" unsur-unsur keislaman dalam materi tes, sehingga peserta tes berkenalan dengan wawasan dunia Islam secara umum. TOAFL lahir dengan visi: "Menjadikan bahasa Arab sebagai bahasa studi Islam dan sains".

TOAFL mulai digunakan sebagai salah satu syarat materi ujian masuk Program S2 dan S3 dan yang akan melanjutkan studi ke luar negeri. Penggunaan TOAFL sebagai materi tes didasarkan pada kebutuhan akademis bahwa para peserta program S2 dan S3 dituntut mampu dan memiliki standar tertentu dalam berbahasa asing, utamanya Arab.

5. Mengukur Butir Tes TOAFL

a) Validitas

Ciri utama tes yang baik adalah kesesuaiannya dengan kemampuan yang diukur, atau yang disebut dengan validitas. Ciri lainnya adalah kemampuannya melakukan pengukuran dengan tingkat keajegan tertentu, yang dapat dikaji menurut beberapa metode. Dengan kata lain, validitas merupakan kesesuaian antara tes dengan apa yang ingin diukur dalam tes itu.

Validitas terbagi beberapa macam, diantaranya adalah validitas isi (*content validity*), validitas konstruk (*construct validity*), dan validitas kriteria (*criteria validity*). Ada juga yang mengklasifikasikan validitas menjadi empat, yaitu: validitas isi, validitas konstruk, validitas prediktif (*predictive validity*), dan validitas konkuren (*concurrent validity*). Validitas isi menuntut adanya kesesuaian isi antara kemampuan yang ingin diukur dan

⁶ J.W. Oller, *Language Test at School: a Pragmatic Approach*, (London: tp. 1979); dan Harold, Madsen, *Technique in Testing*, (Oxford: Oxford University Press, Edisi I, 1983)

tes yang digunakan untuk mengukurnya. Kesesuaian itu tercermin pada jenis kemampuan yang dituntut untuk mengerjakan tes, dibandingkan dengan jenis kemampuan yang dijadikan sasaran pengukuran. Tes dimaksud harus benar-benar memerlukan kemampuan menyimak, dan bukan kemampuan membaca.

Validitas kriteria mengacu kepada kesesuaian antara hasil suatu tes dengan hasil tes lain yang digunakan sebagai kriteria. Kriteria yang digunakan untuk menetapkan tingkat kesesuaian itu dapat diambil dari tes sejenis yang diketahui ciri-cirinya sebagai tes yang baik, dan diselenggarakan pada saat yang hampir bersamaan. Validitas ini juga dikenal sebagai validitas kesetaraan waktu.

Validitas konstruk merupakan sebuah konsep atau teori yang mendasari penggunaan jenis kemampuan, termasuk kemampuan berbahasa. Pembuktiaan adanya validitas konstruk merupakan usaha untuk menunjukkan bahwa skor yang dihasilkan suatu tes benar-benar mencerminkan konstruk yang sama dengan kemampuan yang dijadikan sebagai sasaran pengukurannya. Dalam tes kemampuan *qirâ'ah* (membaca), misalnya, urusan validitas konstruk menyangkut pembuktian apakah skor yang dihasilkan benar-benar mencerminkan jenis dan rincian kemampuan membaca yang sama dengan jenis dan rincian kemampuan yang diperlukan untuk memahami bacaan.⁷

Validitas berkaitan dengan kelayakan penafsiran penggunaan khusus skor hasil tes, sedang validasi merupakan kumpulan bukti untuk menunjukkan dasar saintifik penafsiran skor sebagaimana direncanakan. Gronlund (1985:58) dan Popham (1995:42-43) dalam Burhan Nurgiyanto (2012:153-154) mengemukakan bahwa terdapat tiga pendekatan validitas yang sering dipakai; validitas isi, validitas kriteria dan validitas konstruk. Deskripsinya sesuai tabel berikut:

Tabel 1. Tiga Pendekatan Validasi Tes

Jenis Pendekatan	Prosedur	Makna
Bukti berdasarkan isi (<i>Content-Related Evidence</i>)	Perbandingan butir-butir tes dengan deskripsi spesifikasi tes	Sejauh mana sampel tes mewakili kompetensi ranah yang diukur
Bukti berdasarkan kriteria (<i>Criterion-</i>	Perbandingan skor hasil tes dengan skor kinerja	Sejauh mana kinerja tes mampu memprediksikan

⁷ Djiwandono, M. Soenardi, 1996. *Tes Bahasa dalam Pembelajaran*. Bandung: ITB Pres, hal.17

<i>Related Evidence</i>)	yang kemudian, atau dengan skor kinerja yang sekarang	tampilan kinerja yang akan datang, atau mampu mengestimasi kinerja lain yang dilakukan sekarang
Bukti berdasarkan konstruk (<i>Construct-Related Evidence</i>)	Penetapan makna skor tes dengan mengontrol atau menguji pengembangan tes dan secara eksperimental menentukan berbagai faktor yang mempengaruhi tampilan tes	Seberapa baik kinerja tes dapat ditafsirkan sebagai ukuran yang bermakna dari suatu karakteristik atau kualitas

b) Reliabilitas

Sementara itu, reliabilitas merupakan ciri tes yang memiliki kemampuan untuk menghasilkan pengukuran yang ajeg, tidak berubah-ubah, seandainya digunakan secara berulang-ulang pada sasaran yang sama. Dengan kata lain, reliabilitas terkait bukan dengan tesnya sebagai alat ukur, melainkan dengan hasil pengukurannya dalam bentuk skor yang ajeg. Skor sebagai hasil pengukuran itulah yang seharusnya ajeg, tidak berubah-ubah. Dengan ciri keajegan itu, peserta tes yang sama seharusnya memperoleh skor yang hampir sama pula, seandainya ia kembali mengerjakan tes yang sama, pada kesempatan yang berbeda.

Reliabilitas menunjuk pada konsistensi pengukuran dan bukan ketepatan pengukuran. Tes soal yang reliabel berarti terpercaya atau andal. Jika sebuah tes diujicobakan lebih dari satu kali kepada subyek yang sama dalam waktu yang berbeda dapat menghasilkan data yang kurang lebih sama, tes itu dikatakan reliabel. Alat tes tersebut dapat mengukur secara konsisten, secara ajeg, tidak berubah ubah.

Tingkat reliabilitas (keterandalan) adalah rasio antara skor murni (*true skor*) dengan varian skor yang diperoleh (*observed score*). Dengan kalimat lain, angka yang menunjukkan seberapa banyak variabelitas pada skor yang diperoleh disebabkan oleh perbedaan murni yang ada antara masing-masing individu dalam hal variable yang

diukur. Jika diujikan kepada orang yang sama secara berulang-ulang dan memperlihatkan keajegan, maka TOAFL dinilai reliabel.

c) Daya Beda

Daya beda adalah tingkat yang menunjukkan mampu tidaknya butir tes dalam membedakan antara kemampuan peserta tes yang pandai dan yang kurang pandai. Derajat kesulitan adalah derajat yang menunjukkan sulit tidaknya sebuah butir tes dalam membedakan antara kemampuan peserta tes yang pandai dan yang kurang pandai.

Daya pembeda suatu butir menyatakan seberapa jauh kemampuan butir tersebut mampu membedakan antara kelompok testi (siswa) pandai dengan kelompok testi (siswa) lemah.

Daya pembedaan (D) butir tes dihitung dengan rumus.

$$D = P_H - P_L$$

Keterangan

D = Indeks Daya Pembeda

P_H = Proporsi Siswa kelompok atas yang menjawab benar butir tes

P_L = Proporsi siswa kelompok bawah yang menjawab benar butir tes

Daya pembeda ini sekurang-kurangnya harus berkualitas cukup kriteria yang digunakan untuk menentukan indeks daya pembeda adalah sebagai berikut

Tabel 2. Penafsiran Indeks Daya Pembeda

Indeks Daya Pembeda	Kategori
$0,40 < D$	Butir sangat baik
$0,30 < D \leq 0,40$	Butir baik
$0,20 < D \leq 0,30$	Butir cukup
$D \leq 0,20$	Butir jelek

Dari, data pada Tabel 1 di atas dihitung indeks kesuksesan setiap butir tes, sebagai berikut

Untuk butir 1, $p = 0,88 - 0,25 = 0,63$ (daya pembedaan sangat baik)

Untuk butir 2, $p = 0,63 - 0,38 = 0,25$ (daya pembedaan cukup)

Untuk butir 3, $p = 0,88 - 0,38 = 0,50$ (daya pembedaan sangat baik)

Untuk butir 4, $p = 0,75 - 0,13 = 0,62$ (daya pembedaan sangat baik)

Untuk butir 5, $p = 0,75 - 0,25 = 0,50$ (daya pembedaan sangat baik)

Untuk butir 6, $p = 0,75 - 0,50 = 0,25$ (daya pembedaan cukup)

d) Tingkat Kesukaran

Soal yang ideal adalah soal yang sesuai dengan kemampuan peserta tes. Dengan demikian soal yang terlalu sulit bukan merupakan soal yang baik karena hanya dapat dikerjakan oleh sedikit peserta, khususnya kelompok atas (*upper group*) atau bahkan tidak ada seorangpun yang mampu menyelesaikannya. Demikian pula soal yang terlalu mudah sehingga dapat dikerjakan dengan benar oleh seluruh peserta, sehingga tidak mampu membedakan peserta yang pandai dan yang tidak pandai. Kesimpulannya soal yang baik adalah yang berada pada level sedang. Kondisi soal sebagaimana telah dijelaskan disebut dengan tingkat kesukaran soal (*difficulty level*). Tingkat kesukaran (TK) soal pada dasarnya adalah perbandingan antara banyaknya peserta yang menjawab benar dengan jumlah seluruh peserta, dan dinyatakan dalam bentuk kuantitatif.

Beberapa ahli penilaian memberikan skala tingkat kesukaran (TK) yang berbeda, namun yang banyak dipakai di kalangan pendidikan adalah skala yang dikemukakan oleh Thorndike dan Hagen (*Measurement and Evaluation in Psychology and Education*) sebagai berikut :

Tabel 3. Tingkat Kesukaran

TINGKAT KESUKARAN (TK) atau p	
0,71 - 1,00	: mudah (Md)
0,30 - 0,70	: sedang (Sd)
0,00 - 0,29	: sukar (Sk)

Sedangkan rumus untuk mencari derajat tingkat kesukaran adalah :

$$TK = PB : JP$$

Ket :

TK = Tingkat kesukaran

JP = Jumlah seluruh peserta

PB = Peserta yang menjawab benar

Setelah setiap butir soal dianalisis pada tahap ini, kita akan mengetahui adanya tiga kelompok soal, yaitu soal yang sukar (indeks kurang dari 0,30), soal sedang (indeks 0,30 – 0,70) dan soal mudah (indeks 0,71 – 1,00). Tindak lanjut terhadap tiap-tiap kelompok soal ini adalah :

Tabel 4. Indeks Soal

INDEKS SOAL	TINDAK LANJUT
Sukar (0,00 – 0,29)	- Didrop (tidak digunakan pada tes mendatang) - Dikaji ulang untuk menemukan penyebab soal dianggap sukar (redaksi,pertanyaan, dll) - Digunakan untuk tes tingkat tinggi
Sedang (0,30 – 0,70)	- Dicatat pada bank soal dan dikeluarkan lagi pada tes di waktu yang akan datang
Mudah (0,71 – 1,00)	- Didrop (tidak digunakan pada tes mendatang) - Dikaji ulang untuk menemukan penyebab soal terlalu mudah dijawab - Digunakan untuk tes yang tidak terlalu ketat

C. Hasil Penelitian

Hasil yang diperoleh dari analisis butir soal Maharah Istima' Tes TOAFL UIN Maulana Malik Ibrahim Malang sebagai berikut:

1. Reliabilitas

Temuan pertama dapat diinformasikan bahwa unsur temuan tentang reliabilitas butir tes TOAFL UIN Maulana Malik Ibrahim Malang dapat menjawab rumusan permasalahan pokok pertama dalam penelitian ini. Reliabel artinya dapat diandalkan, dapat dipercaya. Reliabilitas soal dihitung dengan menggunakan rumus KR-20 untuk soal pilihan ganda dan rumus Alpha untuk soal uraian. Hasil penelitian terhadap analisis reliabilitas soal berdasarkan patokan bahwa apabila $r_{11} \geq 0,70$ maka soal yang diujikan memiliki reliabilitas yang tinggi tetapi apabila $r_{11} < 0,70$ maka soal yang diujikan memiliki reliabilitas yang rendah atau tidak reliabel.

Setelah jawaban peserta dari nomor 1 sampai 50 dicocokkan dengan kunci jawaban yang benar lalu dihitung dengan menggunakan rumus tersebut di atas. Setelah melalui perhitungan dengan aplikasi computer program excel diperoleh hasil penelitian bahwa

soal tersebut mempunyai r_{11} lebih besar dari 0,70 yaitu sebesar 0,799 sehingga soal tersebut memiliki reliabilitas yang tinggi atau reliabel.

2. Daya Beda

Kriteria untuk daya pembeda adalah apabila negatif (-) berarti tidak ada daya pembeda, $< 0,20$ dikategorikan daya beda lemah, $0,20 - 0,39$ kategori daya beda cukup, $0,40 - 0,69$ kategori daya beda baik, $0,70 - 1,00$ kategori daya beda baik sekali.

Tabel 5. Daya Beda

NO	DAYA PEMBEDA	KETERANGAN
1.	$0 < 0,20$	Lemah
2.	$0,20 - 0,39$	Cukup
3.	$0,40 - 0,69$	Baik
4.	$0,70 - 1,00$	Baik Sekali

Berdasarkan hasil analisis yang dilakukan menggunakan program SPSS, dapat diketahui bahwa soal dengan daya beda lemah berjumlah 13 soal (26%), soal dengan daya beda cukup berjumlah 9 soal (18%), soal dengan daya beda baik berjumlah 15 soal (30%), soal dengan daya beda baik sekali berjumlah 13 soal (26%).

Tabel 6. Prosentase Daya Beda

NO	DAYA PEMBEDA	KETERANGAN	JUMLAH	PERSENTASE
1.	$0 < 0,20$	Lemah	13	26%
2.	$0,20 - 0,39$	Cukup	9	18%
3.	$0,40 - 0,69$	Baik	15	30%
4.	$0,70 - 1,00$	Baik Sekali	13	26%

3. Tingkat Kesulitan

Tingkat kesulitan soal adalah mengkaji soal-soal tes dari segi kesulitannya sehingga dapat diperoleh soal-soal mana yang termasuk mudah, sedang, dan sulit. Kriteria tingkat kesulitan $0,00 - 0,30$ termasuk soal kategori sulit, $0,31 - 0,70$ termasuk soal kategori sedang, $0,71 - 1,00$ termasuk soal kategori mudah. Berdasarkan hasil penelitian menunjukkan bahwa tidak ada soal yang termasuk kategori sulit ada 5 soal (10%), soal yang termasuk kategori sedang ada 26 soal (52%), dan soal yang termasuk ke dalam kategori mudah ada 19 soal (38%). Hasil

penelitian ini sesuai dengan pendapat dari Anas (2011, 370) yang menyatakan bahwa butir item yang baik apabila butir item tersebut tidak terlalu sukar dan tidak terlalu mudah, dengan kata lain derajat kesulitan item adalah sedang atau cukup. Soal yang mudah membuat siswa tidak ada usaha untuk memecahkannya atau siswa dapat menyepelekan soal. Namun sebaliknya, soal yang sulit membuat siswa putus asa untuk memecahkan soal tersebut. Butir soal yang sulit maupun yang mudah perlu dilakukan tindak lanjut, apakah akan direvisi atau akan dibuang supaya dapat digunakan kembali di ujian yang akan datang. Butir soal yang sedang dapat disimpan di bank soal.

Tabel 7. Tingkat Kesulitan

NO	TINGKAT KESULITAN	JUMLAH	PERSENTASE
1.	0,71 – 1,00 (Mudah)	19	38%
2.	0,31 – 0,70 (Sedang)	26	52%
3.	0,00 – 0,30 (Sulit)	5	10%

Berdasarkan uraian pembahasan diatas dapat kita simpulkan bahwa soal maharah istima' UIN Maulana Malik Ibrahim Malang mempunyai tingkat reliabilitas yang tinggi dan reliabel. Hasil penelitian terhadap analisis reliabilitas soal berdasarkan patokan bahwa apabila $r_{11} \geq 0,70$ maka soal yang diujikan memiliki reliabilitas yang tinggi tetapi apabila $r_{11} < 0,70$ maka soal yang diujikan memiliki reliabilitas yang rendah atau tidak reliabel. Dengan diperolehnya hasil penelitian bahwa soal tersebut mempunyai r_{11} lebih besar dari 0,70 yaitu sebesar 0,799 sehingga soal tersebut memiliki reliabilitas yang tinggi atau reliabel.

Adapun soal maharah istima' TOAFL UIN Maulana Malik Ibrahim Malang secara umum mempunyai daya beda yang cukup baik, yaitu dengan 13 (26%) soal yang mempunyai daya beda yang rendah atau lemah, 9 (18%) soal yang tergolong mempunyai daya beda yang cukup, 15 (30%) soal mempunyai daya beda yang baik, dan 13 (26%) soal yang mempunyai daya beda yang baik sekali. Kriteria untuk daya pembeda adalah apabila negatif (-) berarti tidak ada daya pembeda, $< 0,20$ dikategorikan daya beda lemah, $0,20 - 0,39$ kategori daya beda cukup, $0,40 - 0,69$ kategori daya beda baik, $0,70 - 1,00$ kategori daya beda baik sekali. Sehingga dapat kita

simpulkan bahwa soal maharah istima' TOAFL UIN Maulana Malik Ibrahim Malang mempunyai daya beda yang baik.

Tingkat kesulitan soal adalah mengkaji soal-soal tes dari segi kesulitannya sehingga dapat diperoleh soal-soal mana yang termasuk mudah, sedang, dan sulit. Kriteria tingkat kesulitan 0,00 – 0,30 termasuk soal kategori sulit, 0,31 – 0,70 termasuk soal kategori sedang, 0,71 – 1,00 termasuk soal kategori mudah. Berdasarkan hasil penelitian menunjukkan bahwa tidak ada soal yang termasuk kategori sulit ada 5 soal (10%), soal yang termasuk kategori sedang ada 26 soal (52%), dan soal yang termasuk ke dalam kategori mudah ada 19 soal (38%). Hasil penelitian ini sesuai dengan pendapat dari Anas (2011, 370) yang menyatakan bahwa butir item yang baik apabila butir item tersebut tidak terlalu sukar dan tidak terlalu mudah, dengan kata lain derajat kesulitan item adalah sedang atau cukup. Soal yang mudah membuat siswa tidak ada usaha untuk memecahkannya atau siswa dapat menyepelekan soal. Namun sebaliknya, soal yang sulit membuat siswa putus asa untuk memecahkan soal tersebut. Butir soal yang sulit maupun yang mudah perlu dilakukan tindak lanjut, apakah akan direvisi atau akan dibuang supaya dapat digunakan kembali di ujian yang akan datang. Butir soal yang sedang dapat disimpan di bank soal.

Dengan demikian bahwa soal maharah istima' TOAFL UIN Maulana Malik Ibrahim Malang ada 33 (66%) soal yang direkomendasikan untuk digunakan dan disimpan di bank soal, dan 2 (4%) soal yang disarankan untuk direvisi atau diperbaiki sehingga bisa menjadi soal yang baik dan mempunyai reliabilitas dan daya beda yang baik, kemudian ada 15 (30%) soal yang masih belum mempunyai ciri-ciri atau kriteria soal yang baik, sehingga peneliti menyarankan untuk diganti soal baru yang lebih baik dan mempunyai reliabilitas dan daya beda yang baik.

Tabel 8. Persentase Status Soal

NO	STATUS SOAL	JUMLAH	PERSENTASE
1.	Diterima	33	66%
2.	Direvisi	2	4%
3.	ditolak	15	30%

D. Penutup

Setelah mengetahui tingkat reliabilitas, kesulitan soal, dan daya pembeda, kita dapat menyimpulkan dan menentukan apakah suatu soal dapat dikatakan baik ataukah tidak. Tindak lanjut atas hasil analisis tersebut adalah sebagai berikut :

1. Secara umum soal-soal yang ada pada tes TOAFL UIN Maulana Malik Ibrahim Malang untuk *maharah istima'* dapat dikatakan baik karena telah memadai dan memenuhi syarat dan kriteria soal yang baik.
2. Hasil penelitian terhadap analisis reliabilitas soal tes TOAFL UIN Maulana Malik Ibrahim Malang khususnya *maharah istima'* mempunyai reliabilitas tingkat tinggi, karena berada pada $r_{11} \geq 0,70$ yaitu 0,799, maka soal tersebut memiliki reliabilitas yang tinggi atau reliabel.
3. Berdasarkan hasil analisis yang dilakukan menggunakan program komputer, dapat diketahui bahwa soal dengan daya beda lemah berjumlah 13 soal (26%), soal dengan daya beda cukup berjumlah 9 soal (18%), soal dengan daya beda baik berjumlah 15 soal (30%), soal dengan daya beda baik sekali berjumlah 13 soal (26%). Adapun secara umum soal tersebut mempunyai daya beda yang baik.
4. Berdasarkan hasil penelitian menunjukkan bahwa tidak ada soal yang termasuk kategori sulit ada 5 soal (10%), soal yang termasuk kategori sedang ada 26 soal (52%), dan soal yang termasuk ke dalam kategori mudah ada 19 soal (38%).

Setelah menyelesaikan penelitian ini serta menyimpulkannya, peneliti menyarankan beberapa hal untuk dilaksanakan dan dilanjutkan, yaitu :

- a. Butir soal yang memiliki daya pembeda yang baik dapat langsung dicatat pada bank soal dan dikeluarkan lagi pada tes mendatang, karena kualitasnya memadai.
- b. Butir-butir soal yang daya pembedanya rendah (poor) ada dua kemungkinan :
 - 1) Dikaji ulang dan diperbaiki, selanjutnya dianalisis lagi untuk mengetahui apakah tingkat kesulitan dan daya beda soal semakin baik.
 - 2) Didrop atau dibuang artinya tidak digunakan lagi sebagai soal.
- c. Untuk melanjutkan penelitian di maharah yang lainnya, yaitu kitabah dan qiroah dari soal TOAFL UIN Maulana Malik Ibrahim Malang.