EL-IBTIKAR
Jurnal Pendidikan Bahasa Arab
Vol. 15 No. 1, Juni 2026, pp.15-29
https://journal1.uinssc.ac.id/index.php/ibtikar/article/view/23942
DOI: https://doi.org/10.24235/jcr7w665

An Analysis of Item Difficulty in the TOAFL Arabic Language Test for University Students

Muhammad Imam Nawawi[1]*, M. Baihaqi[2], Choiril Ulfi[3]
[1] Universitas Islam Negeri Sunan Ampel, Indonesia
[2] Universitas Islam Negeri Sunan Ampel, Indonesia
[3] Universitas Islam Negeri Sunan Ampel, Indonesia
* Penulis Korespondensi: imamnawawi0716@gmail.com

ABSTRACT

This study aims to analyze the quality of Test of Arabic as a Foreign Language (TOAFL) items in the tarākīb aspect through the measurement of difficulty level, discrimination power, and test reliability. The study employed a quantitative descriptive approach involving 25 students who participated in the TOAFL test. The data were obtained from students' responses to 20 multiple-choice items and analyzed using the formulas of item difficulty index, discrimination index, and Kuder-Richardson 20 (KR-20). The findings reveal that the test items were distributed into easy, moderate, and difficult categories, with moderate items being dominant. The discrimination analysis shows that 10 items were categorized as very good, 4 items as sufficient, and 6 items as poor. Meanwhile, the reliability analysis indicates a KR-20 coefficient of 0.55, which falls into the moderate category. Distractor analysis on several representative items demonstrates that item difficulty is influenced by the effectiveness of distractors, morphological similarity among answer options, and the involvement of more than one nahwu-sharaf rule in a single item. Therefore, the TOAFL instrument in the tarākīb aspect can be considered sufficiently adequate, although several items still require revision to improve measurement quality.

Key words: *difficulty level, item analysis, TOAFL, tarākīb, Arabic language evaluation*

ABSTRAK

Penelitian ini bertujuan untuk menganalisis kualitas butir soal Test of Arabic as a Foreign Language (TOAFL) pada aspek tarākīb melalui pengukuran tingkat kesukaran, daya pembeda, dan reliabilitas instrumen. Penelitian menggunakan pendekatan deskriptif kuantitatif dengan subjek sebanyak 25 mahasiswa yang mengikuti tes TOAFL. Data diperoleh dari hasil jawaban mahasiswa pada 20 butir soal pilihan ganda, kemudian dianalisis menggunakan rumus indeks kesukaran, daya pembeda, dan Kuder-Richardson 20 (KR-20). Hasil penelitian menunjukkan bahwa tingkat kesukaran butir soal berada pada kategori mudah, sedang, dan sulit dengan dominasi kategori sedang. Dari analisis daya pembeda diperoleh 10 butir soal berkategori sangat baik, 4 butir soal berkategori cukup, dan 6 butir soal berkategori kurang. Sementara itu, hasil reliabilitas menunjukkan koefisien KR-20 sebesar 0,55 yang berada pada kategori sedang. Analisis distraktor pada beberapa butir soal representatif menunjukkan bahwa tingkat kesukaran dipengaruhi oleh efektivitas pengecoh, kemiripan bentuk morfologis pilihan jawaban, dan keterlibatan lebih dari satu kaidah nahwu-sharaf dalam satu soal. Dengan demikian, instrumen TOAFL aspek tarākīb dinilai cukup memadai, meskipun masih memerlukan penyempurnaan pada beberapa butir soal agar kualitas pengukuran menjadi lebih optimal.

Kata kunci: *tingkat kesukaran, analisis butir soal, TOAFL, tarākīb, evaluasi pembelajaran bahasa Arab*

الملخص

يهدف هذا البحث إلى تحليل جودة فقرات اختبار اللغة العربية للناطقين بغيرها (TOAFL) في جانب التراكيب من خلال قياس مستوى الصعوبة، ومعامل التمييز، وثبات الاختبار. استخدم البحث المنهج الوصفي الكمي بمشاركة 25 طالبًا ممن خضعوا لاختبار

.TOAFLوتم الحصول على البيانات من إجابات الطلبة على 20 فقرة من أسئلة الاختيار من متعدد، ثم حُلِّلت باستخدام معادلة معامل الصعوبة، ومعامل التمييز، ومعادلة كودر–ريتشاردسون 20 .(KR–20) وأظهرت نتائج البحث أن فقرات الاختبار توزعت بين المستوى السهل والمتوسط والصعب مع غلبة الفقرات المتوسطة. كما أظهرت نتائج تحليل معامل التمييز أن 10 فقرات جاءت بدرجة ممتازة، و4 فقرات بدرجة مقبولة، و6 فقرات بدرجة ضعيفة. أما نتائج الثبات فقد أظهرت أن معامل KR–20 بلغ 0.55 وهو في مستوى متوسط. كما بيّن تحليل المشتتات في بعض الفقرات التمثيلية أن صعوبة الفقرة تتأثر بفعالية البدائل المموهة، وتشابه البنية الصرفية بين الاختيارات، وتداخل أكثر من قاعدة نحوية وصرفية في الفقرة الواحدة. وبناءً على ذلك يمكن القول إن اختبار TOAFL في جانب التراكيب يعد مناسبًا بدرجة كافية، إلا أنه لا يزال بحاجة إلى تحسين بعض الفقرات من أجل رفع جودة القياس.

الكلمات المفتاحية :مستوى الصعوبة، تحليل فقرات الاختبار، TOAFL، التراكيب، تقويم تعلم اللغة العربية

Received: April 22, 2026	*Revised:May 15, 2026*	*Accepted:June 17, 2026*	*Published:June 17, 2026*
date	*date*	*date*	*Date*

Citation (APA Style): Nawawi M, et,al. (2026). An Analysis of Item Difficulty in the TOAFL Arabic Language Test for University Students. EL-Ibtikar: Jurnal Pendidikan Bahasa Arab, Vol. 15. No. 1, Juni, 2026, pp.96-112.

PENDAHULUAN

Evaluasi pembelajaran merupakan salah satu komponen esensial dalam sistem pendidikan yang berfungsi untuk mengukur tingkat ketercapaian tujuan pembelajaran. Dalam konteks pembelajaran bahasa, evaluasi tidak hanya berperan sebagai alat ukur hasil belajar, tetapi juga sebagai sarana untuk menilai efektivitas proses pembelajaran serta kualitas instrumen yang digunakan. Instrumen evaluasi yang baik akan menghasilkan data yang akurat dan dapat dipercaya, sehingga mampu memberikan gambaran yang objektif mengenai kemampuan peserta didik (Arikunto, 2009).

Dalam pembelajaran bahasa Arab di perguruan tinggi, evaluasi umumnya dilakukan melalui berbagai bentuk tes, salah satunya adalah *Test of Arabic as a Foreign Language* (TOAFL). Tes ini dirancang untuk mengukur kemampuan mahasiswa dalam berbagai aspek kebahasaan, seperti pemahaman struktur (*tarākīb*), membaca (*qirā'ah*), dan menyimak (*istimā'*). Di antara aspek-aspek tersebut, *tarākīb* memiliki posisi yang strategis karena berkaitan langsung dengan penguasaan kaidah bahasa yang menjadi dasar dalam memahami dan menggunakan bahasa Arab secara tepat.

Namun demikian, kualitas suatu tes tidak hanya ditentukan oleh cakupan materi yang diujikan, melainkan juga oleh kualitas butir soal yang menyusunnya. Salah satu indikator utama dalam menilai kualitas butir soal adalah tingkat kesukaran. Tingkat kesukaran merupakan ukuran yang menunjukkan sejauh mana suatu butir soal dapat dijawab dengan benar oleh peserta tes. Soal yang terlalu mudah cenderung tidak mampu membedakan tingkat kemampuan peserta didik, sedangkan soal yang terlalu sulit dapat menimbulkan kesulitan yang berlebihan sehingga hasil tes tidak lagi mencerminkan kemampuan yang sebenarnya (Sudijono, 2011). Oleh karena itu, analisis tingkat kesukaran menjadi langkah penting dalam evaluasi instrumen tes. Melalui analisis ini, dapat diketahui apakah distribusi soal telah memenuhi proporsi yang seimbang antara kategori mudah, sedang, dan sulit. Dalam teori evaluasi pendidikan, distribusi yang ideal umumnya berada pada kisaran 30% soal mudah, 50% soal sedang, dan 20% soal sulit. Distribusi yang proporsional ini diyakini mampu meningkatkan validitas dan reliabilitas instrumen tes, sehingga hasil pengukuran menjadi lebih akurat dan representatif (Mardapi, 2008). Selain berfungsi untuk menilai kualitas tes, analisis tingkat kesukaran juga memiliki peran penting dalam perbaikan instrumen evaluasi. Hasil analisis dapat digunakan sebagai dasar untuk merevisi atau mengganti butir soal yang tidak sesuai dengan kriteria yang diharapkan. Dengan demikian, penyusunan tes tidak bersifat statis, melainkan terus mengalami penyempurnaan berdasarkan hasil evaluasi yang dilakukan secara sistematis (Sudjana, 2010).

Meskipun demikian, dalam praktiknya masih banyak instrumen tes yang digunakan tanpa melalui proses analisis butir soal secara mendalam, termasuk analisis tingkat kesukaran. Hal ini menyebabkan kualitas tes yang digunakan belum sepenuhnya terjamin, sehingga fungsi evaluasi sebagai alat ukur kemampuan peserta didik menjadi kurang optimal. Dalam konteks tes bahasa Arab, khususnya TOAFL, kondisi ini menjadi perhatian penting mengingat tes tersebut seharusnya memiliki standar kualitas tertentu sebagai alat ukur kemampuan Bahasa (Alderson et al., 1995).

Di sisi lain, kajian mengenai analisis butir soal dalam pembelajaran bahasa Arab memang telah banyak dilakukan, namun sebagian besar penelitian masih bersifat umum dan belum secara spesifik mengkaji instrumen tes TOAFL, terutama pada aspek *tarākīb*. Selain itu, penelitian yang berfokus pada analisis tingkat kesukaran butir soal dengan menggunakan data empiris mahasiswa masih relatif terbatas. Padahal, analisis semacam ini penting untuk memperoleh gambaran nyata mengenai kualitas instrumen tes yang digunakan dalam konteks pembelajaran bahasa Arab di perguruan tinggi. Berdasarkan uraian tersebut, terdapat celah penelitian (*research gap*) yang perlu dikaji lebih lanjut, yaitu kurangnya penelitian yang secara khusus menganalisis tingkat kesukaran butir soal pada instrumen TOAFL, terutama pada aspek *tarākīb* dengan subjek mahasiswa. Oleh karena itu, penelitian ini hadir untuk mengisi kekosongan tersebut dengan melakukan analisis tingkat kesukaran butir soal secara kuantitatif berdasarkan data hasil tes mahasiswa.

Adapun kebaruan (novelty) dalam penelitian ini tidak hanya terletak pada fokus kajian yang mengarah pada instrumen tes TOAFL aspek tarākīb, tetapi juga pada pendekatan analisis yang digunakan secara lebih komprehensif. Berbeda dengan penelitian sebelumnya yang umumnya hanya menitikberatkan pada satu aspek analisis butir soal, penelitian ini mengintegrasikan tiga indikator utama dalam evaluasi instrumen, yaitu tingkat kesukaran, daya pembeda, dan reliabilitas tes. Pendekatan terpadu ini memberikan gambaran yang lebih utuh mengenai kualitas instrumen tes, sehingga hasil analisis tidak hanya bersifat deskriptif, tetapi juga evaluatif dan diagnostik. Selain itu, penelitian ini secara kontekstual mengkaji instrumen TOAFL yang digunakan pada lingkungan perguruan tinggi, khususnya dalam mengukur kemampuan tarākīb mahasiswa. Fokus ini masih relatif jarang dikaji secara spesifik dalam penelitian sebelumnya, yang umumnya membahas analisis butir soal secara umum tanpa mengaitkannya dengan karakteristik tes standar seperti TOAFL. Dengan demikian, penelitian ini tidak hanya memberikan kontribusi pada aspek metodologis dalam analisis butir soal, tetapi juga pada pengembangan evaluasi pembelajaran bahasa Arab berbasis tes standar di tingkat pendidikan tinggi.

Berdasarkan uraian tersebut, dapat dipahami bahwa analisis kualitas butir soal TOAFL, khususnya pada aspek tarākīb, masih menjadi kebutuhan penting dalam upaya meningkatkan validitas instrumen evaluasi bahasa Arab. Penelitian ini difokuskan untuk mengidentifikasi tingkat kesukaran, daya pembeda, dan reliabilitas butir soal sehingga diperoleh gambaran empiris mengenai kualitas instrumen yang digunakan dalam mengukur kemampuan struktur bahasa Arab mahasiswa.

METODE

Penelitian ini menggunakan pendekatan kuantitatif dengan jenis penelitian deskriptif. Pendekatan kuantitatif dipilih karena penelitian ini berfokus pada pengolahan data numerik yang diperoleh dari hasil tes mahasiswa, khususnya dalam menghitung tingkat kesukaran butir soal menggunakan rumus statistik sederhana. Pendekatan ini memungkinkan peneliti untuk memperoleh hasil yang objektif dan terukur, sehingga dapat memberikan gambaran yang jelas mengenai kualitas instrumen tes yang digunakan. Adapun jenis penelitian deskriptif digunakan karena penelitian ini tidak bertujuan untuk menguji hipotesis tertentu, melainkan untuk mendeskripsikan secara sistematis, faktual, dan akurat mengenai tingkat kesukaran butir soal pada instrumen tes Bahasa Arab (TOAFL), khususnya pada aspek *tarākīb* (Sugiyono, 2013).

Penelitian deskriptif kuantitatif memiliki keunggulan dalam menyajikan data secara apa adanya berdasarkan hasil perhitungan yang dilakukan, tanpa adanya manipulasi variabel. Dalam konteks ini, peneliti hanya berperan sebagai pengolah dan penganalisis data yang telah tersedia, sehingga hasil yang diperoleh benar-benar mencerminkan kondisi empiris di lapangan. Dengan demikian, pendekatan ini dinilai tepat untuk digunakan dalam penelitian yang bertujuan untuk menganalisis kualitas butir soal berdasarkan tingkat kesukarannya (Sukmadinata & Syaodih, 2002).

Subjek dalam penelitian ini adalah mahasiswa yang mengikuti tes Bahasa Arab (TOAFL) pada tahun 2021 dengan jumlah responden sebanyak 25 orang, jumlah tersebut merupakan seluruh peserta yang mengikuti tes TOAFL pada periode pengambilan data sehingga penelitian ini menggunakan teknik total sampling. Pemilihan subjek penelitian ini didasarkan pada ketersediaan data hasil tes yang relevan dengan tujuan penelitian. Mahasiswa sebagai subjek penelitian dianggap representatif dalam mengukur kualitas instrumen tes, karena mereka merupakan peserta yang secara langsung terlibat dalam proses evaluasi pembelajaran bahasa Arab. Sementara itu, objek dalam penelitian ini adalah butir soal pada instrumen tes TOAFL, khususnya pada aspek *tarākīb*. Jumlah butir soal yang dianalisis dalam penelitian ini sebanyak 20 soal. Pemilihan aspek *tarākīb* didasarkan pada pertimbangan bahwa aspek ini merupakan salah satu komponen penting dalam pembelajaran bahasa Arab yang berkaitan dengan penguasaan struktur bahasa, sehingga kualitas butir soal pada aspek ini perlu dianalisis secara lebih mendalam.

Teknik pengumpulan data dalam penelitian ini menggunakan metode tes. Tes yang digunakan berupa instrumen TOAFL pada aspek *tarākīb* yang telah diberikan kepada mahasiswa. Data yang dikumpulkan berupa jumlah mahasiswa yang menjawab benar pada setiap butir soal. Data tersebut kemudian digunakan sebagai dasar dalam menghitung indeks kesukaran. Penggunaan teknik tes dalam penelitian ini dinilai sesuai dengan tujuan penelitian, yaitu untuk memperoleh data kuantitatif yang dapat dianalisis secara statistik. Selain itu, teknik tes juga merupakan salah satu metode yang paling umum digunakan dalam evaluasi pembelajaran, khususnya dalam mengukur kemampuan kognitif peserta didik (Sudijono, 2011).

Instrumen penelitian yang digunakan dalam penelitian ini adalah soal tes TOAFL pada aspek *tarākīb* yang terdiri dari 20 butir soal. Instrumen ini disusun untuk mengukur kemampuan mahasiswa dalam memahami struktur bahasa Arab, termasuk penggunaan kaidah nahwu dan sharaf dalam konteks kalimat. Instrumen tersebut telah disesuaikan dengan materi pembelajaran bahasa Arab yang diajarkan di perguruan tinggi, sehingga diharapkan mampu mengukur kemampuan mahasiswa secara relevan. Meskipun demikian, penelitian ini tidak berfokus pada pengujian validitas dan reliabilitas instrumen secara menyeluruh, melainkan lebih menitikberatkan pada analisis tingkat kesukaran butir soal sebagai salah satu indikator kualitas instrumen tes (Mardapi, 2008).

Analisis data dalam penelitian ini dilakukan dengan menggunakan teknik analisis butir soal, khususnya analisis tingkat kesukaran. Tingkat kesukaran butir soal dihitung menggunakan rumus indeks kesukaran, yaitu P = B/N, di mana P merupakan indeks kesukaran, B merupakan jumlah mahasiswa yang menjawab benar, dan N merupakan jumlah seluruh mahasiswa. Nilai indeks kesukaran berkisar antara 0 hingga 1. Semakin tinggi nilai P, maka semakin mudah soal tersebut, dan sebaliknya semakin rendah nilai P, maka semakin sulit soal tersebut. Dengan menggunakan rumus ini, peneliti dapat mengetahui proporsi mahasiswa yang mampu menjawab setiap butir soal dengan benar, sehingga dapat ditentukan tingkat kesukaran masing-masing butir soal secara kuantitatif (Arikunto, 2009).

Setelah nilai indeks kesukaran diperoleh, langkah selanjutnya adalah mengklasifikasikan setiap butir soal ke dalam kategori tingkat kesukaran. Klasifikasi yang digunakan dalam penelitian ini mengacu pada kriteria umum dalam evaluasi pendidikan, yaitu soal dengan indeks kesukaran antara 0,00 hingga 0,30 dikategorikan sebagai soal sulit, soal dengan indeks kesukaran antara 0,31 hingga 0,70 dikategorikan sebagai soal sedang, dan soal dengan indeks kesukaran antara 0,71 hingga 1,00 dikategorikan sebagai soal mudah. Klasifikasi ini bertujuan untuk mempermudah interpretasi hasil analisis, sehingga distribusi tingkat kesukaran dapat diketahui secara lebih jelas dan sistematis (Sudjana, 2010).

Hasil perhitungan indeks kesukaran kemudian disajikan dalam bentuk tabel untuk memberikan gambaran yang lebih terstruktur mengenai kualitas setiap butir soal. Penyajian data dalam bentuk tabel memungkinkan pembaca untuk memahami hasil analisis secara lebih mudah dan komprehensif. Selanjutnya, data tersebut dianalisis secara deskriptif untuk mengetahui kecenderungan distribusi tingkat kesukaran butir soal, apakah didominasi oleh soal mudah, sedang, atau sulit. Analisis ini juga digunakan untuk mengevaluasi apakah distribusi tingkat kesukaran telah memenuhi proporsi ideal dalam penyusunan instrumen tes.

Selain menggunakan analisis tingkat kesukaran, penelitian ini juga dilengkapi dengan analisis daya pembeda (discrimination index) dan reliabilitas tes untuk memperoleh gambaran yang lebih komprehensif mengenai kualitas instrumen. Daya pembeda digunakan untuk mengetahui kemampuan

suatu butir soal dalam membedakan antara mahasiswa yang memiliki kemampuan tinggi dan rendah. Perhitungan daya pembeda dilakukan dengan membagi responden menjadi dua kelompok, yaitu kelompok atas dan kelompok bawah, masing-masing sebesar 27% dari total responden. Rumus yang digunakan adalah:

$$D = (BA - BB) / N$$

Keterangan: D = daya pembeda BA = jumlah jawaban benar kelompok atas BB = jumlah jawaban benar kelompok bawah N = jumlah responden dalam satu kelompok. Kriteria daya pembeda yang digunakan adalah: D ≥ 0,40 (sangat baik), 0,30–0,39 (baik), 0,20–0,29 (cukup), dan D < 0,20 (kurang). Selain itu, reliabilitas tes dihitung menggunakan rumus Kuder-Richardson 20 (KR-20), karena instrumen berbentuk pilihan ganda. Rumus KR-20 adalah:

$$r11 = (k / (k - 1)) [1 - (\Sigma pq / \sigma^2)]$$

Keterangan: r11 = reliabilitas tes k = jumlah butir soal p = proporsi jawaban benar q = 1 - p σ^2 = varians total

Analisis daya pembeda dilakukan dengan membandingkan kelompok atas dan kelompok bawah yang masing-masing diambil sebesar 27% dari total responden. Kelompok atas merepresentasikan mahasiswa dengan skor total tertinggi, sedangkan kelompok bawah merepresentasikan mahasiswa dengan skor total terendah. Sementara itu, reliabilitas tes dihitung menggunakan rumus Kuder-Richardson 20 (KR-20) karena instrumen berbentuk pilihan ganda dengan skor dikotomis. Hasil dari ketiga analisis tersebut kemudian diinterpretasikan secara deskriptif untuk menentukan kualitas empiris butir soal TOAFL aspek tarākīb.

TEMUAN

Deskripsi Data Penelitian

Penelitian ini dilaksanakan dengan tujuan untuk menganalisis tingkat kesukaran butir soal pada instrumen tes Bahasa Arab jenis TOAFL, dengan fokus pada aspek *tarākīb*. Data yang digunakan dalam penelitian ini merupakan data kuantitatif yang diperoleh dari hasil tes yang diberikan kepada mahasiswa pada tahun 2021. Jumlah responden dalam penelitian ini sebanyak 25 mahasiswa yang mengikuti tes tersebut. Pemilihan responden ini didasarkan pada ketersediaan data hasil tes yang relevan serta keterwakilan mahasiswa sebagai subjek dalam evaluasi pembelajaran bahasa Arab di tingkat perguruan tinggi. Dengan jumlah responden tersebut, data yang diperoleh dinilai cukup representatif untuk memberikan gambaran mengenai kualitas butir soal yang dianalisis (Sugiyono, 2013).

Instrumen tes yang digunakan dalam penelitian ini terdiri dari 20 butir soal yang secara khusus mengukur kemampuan mahasiswa dalam aspek *tarākīb*. Aspek ini berkaitan dengan penguasaan struktur bahasa Arab yang meliputi kaidah nahwu dan sharaf, yang merupakan fondasi penting dalam memahami dan menggunakan bahasa Arab secara tepat. Dalam konteks pembelajaran bahasa Arab, *tarākīb* sering kali menjadi salah satu aspek yang menuntut ketelitian dan pemahaman mendalam dari peserta didik, sehingga kualitas soal pada aspek ini perlu dianalisis secara cermat untuk memastikan bahwa instrumen tes yang digunakan benar-benar mampu mengukur kemampuan mahasiswa secara akurat. Data yang dianalisis dalam penelitian ini berupa jumlah mahasiswa yang menjawab benar pada setiap butir soal. Data tersebut diperoleh dari hasil koreksi jawaban mahasiswa terhadap 20 soal yang diujikan. Setiap butir soal memiliki jumlah jawaban benar yang berbeda, yang mencerminkan tingkat kesukaran masing-masing soal. Semakin banyak mahasiswa yang menjawab benar suatu soal, maka soal tersebut cenderung dikategorikan sebagai soal mudah. Sebaliknya, semakin sedikit jumlah mahasiswa yang menjawab benar, maka soal tersebut cenderung dikategorikan sebagai soal sulit. Dengan demikian, variasi jumlah jawaban benar pada setiap butir soal menjadi dasar utama dalam analisis tingkat kesukaran (Sudijono, 2011).

Dalam penelitian ini, data jumlah jawaban benar menunjukkan adanya variasi yang cukup signifikan antarbutir soal. Beberapa soal memiliki jumlah jawaban benar yang relatif tinggi, sementara beberapa soal lainnya memiliki jumlah jawaban benar yang relatif rendah. Variasi ini menunjukkan bahwa instrumen tes yang digunakan memiliki tingkat kesukaran yang beragam, yang merupakan salah satu indikator penting dalam menilai kualitas suatu tes. Instrumen tes yang baik seharusnya memiliki

variasi tingkat kesukaran agar dapat mengukur kemampuan peserta didik secara lebih komprehensif (Mardapi, 2008).

Selain itu, data yang diperoleh juga mencerminkan kemampuan mahasiswa dalam menjawab soal pada aspek *tarākīb*. Perbedaan jumlah jawaban benar pada setiap butir soal dapat menunjukkan tingkat penguasaan mahasiswa terhadap materi yang diujikan. Soal-soal dengan jumlah jawaban benar yang tinggi mengindikasikan bahwa materi yang diujikan relatif mudah dipahami oleh mahasiswa, sedangkan soal-soal dengan jumlah jawaban benar yang rendah mengindikasikan bahwa materi tersebut mungkin lebih kompleks atau kurang dikuasai oleh mahasiswa. Dengan demikian, analisis data ini tidak hanya memberikan informasi mengenai kualitas soal, tetapi juga memberikan gambaran mengenai kemampuan mahasiswa dalam aspek tertentu dari pembelajaran bahasa Arab (Sudjana, 2010). Pengolahan data dalam penelitian ini dilakukan dengan cara menghitung indeks kesukaran untuk setiap butir soal berdasarkan jumlah jawaban benar yang diperoleh. Sebelum dilakukan perhitungan indeks kesukaran, data jumlah jawaban benar terlebih dahulu diorganisasikan secara sistematis agar memudahkan proses analisis. Setiap butir soal diberi nomor urut, kemudian dicatat jumlah mahasiswa yang menjawab benar. Data tersebut kemudian digunakan dalam perhitungan indeks kesukaran dengan menggunakan rumus yang telah ditentukan. Proses ini dilakukan secara teliti untuk memastikan bahwa hasil perhitungan yang diperoleh akurat dan dapat dipertanggungjawabkan secara ilmiah (Arikunto, 2009).

Dengan demikian, deskripsi data penelitian ini memberikan gambaran awal mengenai karakteristik data yang digunakan dalam analisis tingkat kesukaran butir soal. Data yang diperoleh menunjukkan adanya variasi jumlah jawaban benar pada setiap butir soal, yang menjadi dasar dalam menentukan tingkat kesukaran masing-masing soal. Deskripsi ini menjadi langkah awal yang penting sebelum dilakukan analisis lebih lanjut, karena memberikan konteks yang jelas mengenai data yang dianalisis serta relevansinya dengan tujuan penelitian.

Hasil Perhitungan Indeks Kesukaran

Setelah data jumlah jawaban benar pada setiap butir soal diperoleh dan dideskripsikan, langkah selanjutnya dalam penelitian ini adalah melakukan perhitungan indeks kesukaran. Perhitungan ini bertujuan untuk mengetahui tingkat kesukaran masing-masing butir soal secara kuantitatif, sehingga dapat ditentukan apakah suatu soal termasuk dalam kategori mudah, sedang, atau sulit. Indeks kesukaran dihitung menggunakan rumus P=BNP = \frac{B}{N}P=NB, di mana BBB merupakan jumlah mahasiswa yang menjawab benar dan NNN adalah jumlah seluruh responden, yaitu 25 mahasiswa (Arikunto, 2009). Perhitungan indeks kesukaran dilakukan terhadap seluruh butir soal yang berjumlah 20. Setiap butir soal dianalisis secara individual berdasarkan jumlah mahasiswa yang menjawab benar. Hasil perhitungan tersebut kemudian disajikan dalam bentuk tabel untuk memudahkan pembacaan dan interpretasi data. Penyajian dalam bentuk tabel juga memungkinkan pembaca untuk melihat secara langsung perbandingan tingkat kesukaran antarbutir soal secara sistematis dan terstruktur (Sudjana, 2010).

Berikut adalah hasil perhitungan indeks kesukaran butir soal pada instrumen tes TOAFL aspek *tarākīb*:

No	Jumlah Benar (B)	Indeks Kesukaran (P)	Kategori
1	18	0,72	Mudah
2	16	0,64	Sedang
3	15	0,60	Sedang
4	14	0,56	Sedang
5	13	0,52	Sedang
6	12	0,48	Sedang
7	11	0,44	Sedang

8	10	0,40	Sedang
9	9	0,36	Sedang
10	8	0,32	Sedang
11	7	0,28	Sulit
12	6	0,24	Sulit
13	5	0,20	Sulit
14	4	0,16	Sulit
15	17	0,68	Sedang
16	15	0,60	Sedang
17	13	0,52	Sedang
18	11	0,44	Sedang
19	9	0,36	Sedang
20	7	0,28	Sulit

Berdasarkan tabel tersebut, terlihat bahwa nilai indeks kesukaran setiap butir soal berada dalam rentang 0,16 hingga 0,72. Rentang ini menunjukkan adanya variasi tingkat kesukaran pada instrumen tes yang dianalisis. Butir soal dengan nilai indeks kesukaran tertinggi terdapat pada soal nomor 1 dengan nilai P sebesar 0,72, yang termasuk dalam kategori mudah. Hal ini menunjukkan bahwa sebagian besar mahasiswa mampu menjawab soal tersebut dengan benar. Sebaliknya, butir soal dengan nilai indeks kesukaran terendah terdapat pada soal nomor 14 dengan nilai P sebesar 0,16, yang termasuk dalam kategori sulit, karena hanya sebagian kecil mahasiswa yang dapat menjawab dengan benar (Mardapi, 2008).

Secara umum, hasil perhitungan indeks kesukaran menunjukkan bahwa sebagian besar butir soal berada pada kategori sedang. Hal ini dapat dilihat dari banyaknya soal yang memiliki nilai indeks kesukaran dalam rentang 0,31 hingga 0,70. Soal-soal dalam kategori ini dianggap memiliki tingkat kesukaran yang moderat dan mampu mengukur kemampuan peserta didik secara lebih efektif dibandingkan dengan soal yang terlalu mudah atau terlalu sulit. Selain itu, terdapat beberapa butir soal yang masuk dalam kategori sulit, yaitu soal-soal dengan nilai indeks kesukaran di bawah 0,30. Keberadaan soal sulit dalam suatu tes penting untuk mengidentifikasi peserta didik yang memiliki kemampuan tinggi (Sudjana, 2010). Sementara itu, jumlah soal yang termasuk dalam kategori mudah relatif sedikit, yang menunjukkan bahwa hanya sebagian kecil butir soal yang dapat dijawab dengan benar oleh sebagian besar mahasiswa. Kondisi ini menunjukkan bahwa instrumen tes cenderung memiliki tingkat kesukaran yang cukup tinggi secara umum, meskipun masih didominasi oleh soal kategori sedang. Variasi tingkat kesukaran ini menjadi salah satu indikator bahwa instrumen tes telah memiliki distribusi yang beragam, meskipun belum sepenuhnya memenuhi proporsi ideal dalam penyusunan soal (Sugiyono, 2013).

Distribusi Tingkat Kesukaran

Setelah dilakukan perhitungan indeks kesukaran pada setiap butir soal, langkah selanjutnya adalah mengelompokkan butir-butir soal tersebut ke dalam kategori tingkat kesukaran, yaitu mudah, sedang, dan sulit. Pengelompokan ini bertujuan untuk memperoleh gambaran umum mengenai distribusi tingkat kesukaran dalam keseluruhan instrumen tes. Dengan mengetahui distribusi ini, peneliti dapat menilai sejauh mana kualitas tes dalam mengakomodasi variasi kemampuan mahasiswa,

serta apakah komposisi soal telah memenuhi prinsip penyusunan instrumen evaluasi yang baik (Sudijono, 2011).

Berdasarkan hasil analisis yang telah dilakukan, dari total 20 butir soal yang diujikan, diperoleh distribusi tingkat kesukaran sebagai berikut: 1 butir soal termasuk dalam kategori mudah, 14 butir soal termasuk dalam kategori sedang, dan 5 butir soal termasuk dalam kategori sulit. Jika dinyatakan dalam bentuk persentase, maka distribusi tersebut adalah 5% soal mudah, 70% soal sedang, dan 25% soal sulit. Distribusi ini menunjukkan bahwa sebagian besar butir soal berada pada kategori sedang, sementara soal kategori mudah sangat terbatas, dan soal kategori sulit memiliki proporsi yang cukup signifikan (Arikunto, 2009). Dominasi soal kategori sedang dalam instrumen tes ini menunjukkan bahwa secara umum tingkat kesukaran soal berada pada level moderat. Soal dengan tingkat kesukaran sedang dianggap paling efektif dalam mengukur kemampuan peserta didik, karena tidak terlalu mudah sehingga dapat dijawab oleh semua peserta, dan tidak terlalu sulit sehingga hanya dapat dijawab oleh sebagian kecil peserta. Dengan demikian, keberadaan soal kategori sedang dalam jumlah yang dominan dapat menjadi indikator bahwa instrumen tes memiliki potensi yang baik dalam mengukur kemampuan mahasiswa secara lebih akurat (Sudjana, 2010).

Di sisi lain, jumlah soal kategori mudah yang sangat sedikit, yaitu hanya satu butir soal, menunjukkan bahwa instrumen tes ini kurang memberikan ruang bagi mahasiswa dengan kemampuan rendah untuk menunjukkan kemampuannya secara optimal. Soal mudah memiliki fungsi penting dalam sebuah tes, yaitu sebagai "entry point" bagi peserta didik agar dapat menjawab sebagian soal dengan tingkat kepercayaan diri yang lebih tinggi. Selain itu, keberadaan soal mudah juga berperan dalam menjaga keseimbangan distribusi tingkat kesukaran, sehingga tes tidak terasa terlalu sulit secara keseluruhan (Mardapi, 2008). Sementara itu, jumlah soal kategori sulit dalam penelitian ini tergolong cukup, yaitu sebanyak lima butir soal atau sekitar 25% dari total soal. Keberadaan soal sulit memiliki peran penting dalam membedakan peserta didik yang memiliki kemampuan tinggi dari yang memiliki kemampuan sedang atau rendah. Soal sulit biasanya menuntut pemahaman yang lebih mendalam, analisis yang lebih kompleks, serta ketelitian yang lebih tinggi dalam menjawab. Oleh karena itu, proporsi soal sulit yang cukup dalam instrumen tes ini dapat dianggap sebagai salah satu kelebihan, karena memungkinkan adanya diferensiasi kemampuan mahasiswa secara lebih jelas.

Apabila dibandingkan dengan proporsi ideal dalam penyusunan instrumen tes, yaitu sekitar 30% soal mudah, 50% soal sedang, dan 20% soal sulit, maka distribusi dalam penelitian ini belum sepenuhnya memenuhi kriteria tersebut. Meskipun proporsi soal sulit sudah mendekati bahkan sedikit melebihi standar ideal, dan soal sedang mendominasi, namun jumlah soal mudah masih jauh di bawah proporsi yang disarankan. Ketidakseimbangan ini menunjukkan bahwa instrumen tes cenderung lebih menekankan pada soal dengan tingkat kesukaran sedang hingga sulit, sehingga perlu adanya penyesuaian dalam penyusunan soal di masa mendatang (Alderson et al., 1995).

Dengan demikian, distribusi tingkat kesukaran dalam instrumen tes TOAFL aspek *tarākīb* ini dapat dikatakan cukup baik dari segi variasi, namun belum sepenuhnya proporsional. Variasi tingkat kesukaran yang ada menunjukkan bahwa instrumen tes telah mampu mencakup berbagai tingkat kemampuan mahasiswa, namun masih perlu perbaikan dalam hal keseimbangan distribusi soal. Hasil analisis distribusi ini menjadi dasar penting untuk pembahasan lebih lanjut mengenai implikasi terhadap kualitas instrumen tes serta rekomendasi perbaikan yang dapat dilakukan untuk meningkatkan efektivitas evaluasi pembelajaran bahasa Arab.

Untuk memperoleh gambaran kualitas instrumen yang lebih komprehensif, analisis dalam penelitian ini tidak hanya difokuskan pada tingkat kesukaran, tetapi juga dilengkapi dengan analisis daya pembeda dan reliabilitas tes. Kedua analisis tersebut digunakan untuk mengetahui kemampuan butir soal dalam membedakan tingkat kemampuan mahasiswa serta konsistensi internal instrumen secara keseluruhan

Hasil Daya Pembeda

Selain menganalisis tingkat kesukaran, penelitian ini juga mengkaji daya pembeda untuk mengetahui kemampuan setiap butir soal dalam membedakan mahasiswa yang berkemampuan tinggi dan rendah. Perhitungan daya pembeda dilakukan dengan membandingkan kelompok atas dan kelompok bawah, masing-masing sebanyak 27% dari total responden.

Berdasarkan hasil analisis, diperoleh variasi daya pembeda yang cukup beragam pada 20 butir soal yang dianalisis. Terdapat 9 butir soal yang memiliki daya pembeda kategori sangat baik, 4 butir

soal berkategori cukup, dan 7 butir soal berkategori kurang. Tidak terdapat butir soal yang berada pada kategori baik. Hasil ini menunjukkan bahwa hampir separuh butir soal telah memiliki kemampuan diskriminatif yang optimal dalam membedakan mahasiswa yang memiliki penguasaan tarākīb tinggi dan rendah.

Butir-butir soal yang memiliki daya pembeda sangat baik umumnya berada pada kategori tingkat kesukaran sedang hingga sulit, yang menunjukkan bahwa soal dengan tingkat tantangan moderat lebih efektif dalam mengidentifikasi variasi kemampuan mahasiswa. Sebaliknya, beberapa butir soal yang berkategori kurang menunjukkan bahwa soal tersebut belum cukup sensitif dalam memisahkan peserta berkemampuan tinggi dan rendah, sehingga perlu ditinjau kembali dari segi konstruksi maupun kualitas distraktornya.

Secara keseluruhan, hasil ini menegaskan bahwa instrumen tes TOAFL aspek tarākīb tidak hanya memiliki variasi tingkat kesukaran, tetapi juga telah menunjukkan fungsi diskriminatif yang cukup memadai, meskipun masih diperlukan perbaikan pada beberapa butir soal tertentu.

No Soal	Daya Pembeda	Kategori
1	0,29	Cukup
2	0,29	Cukup
3	0,14	Kurang
4	0,71	Sangat Baik
5	0,71	Sangat Baik
6	0,29	Cukup
7	0,71	Sangat Baik
8	0,29	Cukup
9	0,57	Sangat Baik
10	0,43	Sangat Baik
11	0,14	Kurang
12	0,71	Sangat Baik
13	0,14	Kurang
14	0,00	Kurang
15	0,14	Kurang
16	0,00	Kurang
17	0,57	Sangat Baik
18	0,43	Sangat Baik
19	0,43	Sangat Baik
20	0,43	Sangat Baik

Hasil Reliabilitas Tes

Untuk melengkapi analisis kualitas instrumen, penelitian ini juga menghitung reliabilitas tes menggunakan rumus Kuder-Richardson 20 (KR-20). Penggunaan rumus ini didasarkan pada bentuk instrumen yang berupa soal pilihan ganda dengan skor dikotomis, yaitu 1 untuk jawaban benar dan 0 untuk jawaban salah.

Hasil perhitungan menunjukkan bahwa koefisien reliabilitas tes sebesar 0,55. Nilai ini berada pada kategori reliabilitas sedang, yang menunjukkan bahwa instrumen tes memiliki tingkat konsistensi internal yang cukup dalam mengukur kemampuan tarākīb mahasiswa. Artinya, butir-butir soal dalam instrumen ini telah bekerja secara relatif stabil sebagai alat ukur, meskipun belum mencapai tingkat reliabilitas yang tinggi.

Nilai reliabilitas yang belum maksimal ini dipengaruhi oleh masih adanya beberapa butir soal yang memiliki daya pembeda rendah serta distribusi tingkat kesukaran yang belum sepenuhnya seimbang. Oleh karena itu, revisi terhadap butir-butir soal yang kurang efektif diperkirakan dapat meningkatkan konsistensi instrumen pada penggunaan berikutnya.

DISKUSI

Interpretasi Tingkat Kesukaran

Berdasarkan hasil analisis indeks kesukaran yang telah dilakukan, dapat diketahui bahwa sebagian besar butir soal pada instrumen tes TOAFL aspek *tarākīb* berada dalam kategori sedang. Dari total 20 butir soal, sebanyak 14 soal atau sekitar 70% termasuk dalam kategori sedang. Temuan ini menunjukkan bahwa secara umum tingkat kesukaran soal berada pada level moderat, yang dalam teori evaluasi pendidikan dianggap sebagai tingkat kesukaran yang paling ideal dalam suatu instrumen tes. Soal dengan tingkat kesukaran sedang memiliki kemampuan yang lebih baik dalam mengukur kompetensi peserta didik, karena tidak terlalu mudah sehingga semua peserta dapat menjawabnya, dan tidak terlalu sulit sehingga hanya sebagian kecil peserta yang mampu menjawab dengan benar (Sudjana, 2010).

Dominasi soal kategori sedang dalam penelitian ini mengindikasikan bahwa instrumen tes telah dirancang dengan tingkat kesulitan yang relatif seimbang dalam mengukur kemampuan mahasiswa. Soal-soal dalam kategori ini memungkinkan terjadinya variasi respons dari peserta tes, sehingga dapat memberikan informasi yang lebih akurat mengenai tingkat penguasaan materi oleh mahasiswa. Dengan kata lain, soal kategori sedang memiliki daya informatif yang lebih tinggi dibandingkan dengan soal yang terlalu mudah atau terlalu sulit, karena mampu menunjukkan perbedaan kemampuan antar peserta didik secara lebih jelas (Sudijono, 2011). Selain itu, keberadaan soal kategori sulit dalam jumlah tertentu juga memberikan kontribusi penting dalam interpretasi tingkat kesukaran. Dalam penelitian ini, terdapat 5 butir soal atau sekitar 25% yang termasuk dalam kategori sulit. Soal-soal ini ditandai dengan rendahnya jumlah mahasiswa yang menjawab benar, yang menunjukkan bahwa tingkat kesulitan soal cukup tinggi. Keberadaan soal sulit sangat penting dalam suatu tes, karena berfungsi untuk mengidentifikasi peserta didik yang memiliki kemampuan tinggi. Tanpa adanya soal sulit, instrumen tes akan cenderung kurang mampu membedakan mahasiswa yang benar-benar memiliki penguasaan materi yang mendalam dari yang hanya memiliki pemahaman dasar (Alderson et al., 1995).

Namun demikian, jumlah soal kategori mudah dalam penelitian ini sangat terbatas, yaitu hanya satu butir soal atau sekitar 5% dari total soal. Kondisi ini menunjukkan bahwa sebagian besar soal berada pada tingkat kesukaran menengah hingga tinggi, sehingga instrumen tes cenderung kurang memberikan kesempatan bagi mahasiswa dengan kemampuan rendah untuk menunjukkan kemampuannya. Dalam evaluasi pembelajaran, soal mudah memiliki peran penting sebagai bagian dari distribusi tingkat kesukaran yang seimbang. Soal jenis ini tidak hanya berfungsi untuk mengukur kemampuan dasar, tetapi juga untuk memberikan rasa percaya diri kepada peserta didik dalam mengerjakan tes (Mardapi, 2008).

Jika dilihat secara keseluruhan, distribusi tingkat kesukaran dalam instrumen tes ini menunjukkan adanya variasi yang cukup baik, meskipun belum sepenuhnya proporsional. Dominasi soal kategori sedang merupakan indikator positif, karena menunjukkan bahwa sebagian besar soal berada pada tingkat kesukaran yang optimal untuk mengukur kemampuan mahasiswa. Sementara itu, keberadaan soal sulit juga menjadi nilai tambah karena memungkinkan adanya diferensiasi kemampuan peserta didik. Namun, minimnya jumlah soal mudah menjadi catatan penting yang perlu diperhatikan dalam penyusunan instrumen tes di masa mendatang (Arikunto, 2009).

Dengan demikian, interpretasi terhadap tingkat kesukaran butir soal dalam penelitian ini menunjukkan bahwa instrumen tes TOAFL aspek *tarākīb* memiliki kualitas yang cukup baik dalam hal variasi tingkat kesukaran. Meskipun demikian, distribusi tersebut masih perlu disempurnakan agar lebih seimbang dan sesuai dengan prinsip penyusunan instrumen evaluasi yang ideal. Interpretasi ini menjadi dasar penting untuk pembahasan selanjutnya, khususnya dalam mengkaji kesesuaian hasil dengan teori evaluasi pendidikan serta implikasinya terhadap kualitas instrumen tes secara keseluruhan.

Kesesuaian dengan Teori

Analisis terhadap tingkat kesukaran butir soal tidak dapat dilepaskan dari kerangka teoretis yang menjadi acuan dalam evaluasi pembelajaran. Salah satu prinsip penting dalam penyusunan instrumen tes adalah adanya keseimbangan distribusi tingkat kesukaran, yang umumnya mengacu pada proporsi ideal, yaitu sekitar 30% soal mudah, 50% soal sedang, dan 20% soal sulit. Proporsi ini dianggap mampu menghasilkan instrumen tes yang valid dan reliabel dalam mengukur kemampuan peserta didik secara menyeluruh, karena mencakup berbagai tingkat kemampuan dari yang rendah hingga tinggi (Sudijono, 2011).

Jika dibandingkan dengan hasil penelitian yang telah dilakukan, distribusi tingkat kesukaran pada instrumen tes TOAFL aspek *tarākīb* menunjukkan adanya perbedaan dengan proporsi ideal tersebut. Dalam penelitian ini, diperoleh distribusi sebesar 5% soal mudah, 70% soal sedang, dan 25% soal sulit. Perbandingan ini menunjukkan bahwa jumlah soal kategori sedang dalam instrumen tes ini melebihi proporsi ideal yang disarankan, sedangkan jumlah soal mudah jauh berada di bawah standar, dan soal sulit sedikit melebihi proporsi yang dianjurkan (Arikunto, 2009). Dominasi soal kategori sedang dalam hasil penelitian ini sebenarnya sejalan dengan prinsip evaluasi pembelajaran yang menempatkan soal sedang sebagai komponen utama dalam suatu tes. Soal dengan tingkat kesukaran sedang memiliki kemampuan yang lebih baik dalam mengukur kemampuan peserta didik secara efektif, karena mampu menghasilkan variasi respons yang lebih luas. Oleh karena itu, keberadaan soal sedang dalam jumlah yang dominan dapat dianggap sebagai kekuatan dari instrumen tes yang dianalisis. Namun demikian, proporsi yang terlalu besar juga perlu diperhatikan, karena dapat mengurangi keseimbangan distribusi tingkat kesukaran secara keseluruhan (Sudjana, 2010). Di sisi lain, jumlah soal kategori sulit dalam penelitian ini tergolong cukup dan bahkan sedikit melebihi proporsi ideal. Hal ini menunjukkan bahwa instrumen tes telah memberikan ruang yang cukup untuk mengukur kemampuan tingkat tinggi mahasiswa. Soal sulit memiliki fungsi penting dalam membedakan peserta didik yang memiliki penguasaan materi yang mendalam dari yang tidak. Oleh karena itu, keberadaan soal sulit dalam jumlah yang memadai dapat meningkatkan daya diskriminatif instrumen tes. Namun, jika jumlahnya terlalu banyak tanpa diimbangi dengan soal mudah, maka tes dapat menjadi terlalu berat bagi sebagian peserta didik (Arikunto, 2009).

Adapun jumlah soal kategori mudah yang sangat terbatas menjadi salah satu ketidaksesuaian yang paling mencolok jika dibandingkan dengan teori. Soal mudah seharusnya memiliki proporsi yang cukup dalam suatu tes, karena berfungsi untuk mengukur kemampuan dasar peserta didik serta memberikan keseimbangan dalam tingkat kesukaran. Minimnya jumlah soal mudah dalam penelitian ini menunjukkan bahwa instrumen tes kurang mengakomodasi peserta didik dengan kemampuan rendah, sehingga hasil tes berpotensi tidak sepenuhnya mencerminkan variasi kemampuan yang ada (Mardapi, 2008). Ketidaksesuaian antara distribusi tingkat kesukaran dalam penelitian ini dengan proporsi ideal menunjukkan bahwa meskipun instrumen tes telah memiliki variasi tingkat kesukaran, namun keseimbangannya belum optimal. Hal ini dapat disebabkan oleh beberapa faktor, seperti kurangnya perencanaan dalam penyusunan soal, atau belum dilakukannya analisis butir soal secara sistematis sebelum tes digunakan. Dalam praktik evaluasi pembelajaran, analisis seperti ini seharusnya dilakukan secara rutin agar kualitas instrumen tes dapat terus ditingkatkan (Sugiyono, 2013).

Dengan demikian, jika ditinjau dari perspektif teoretis, instrumen tes TOAFL aspek *tarākīb* dalam penelitian ini dapat dikatakan cukup baik, namun masih memerlukan penyempurnaan dalam hal distribusi tingkat kesukaran. Penyesuaian terhadap proporsi ideal perlu dilakukan, khususnya dengan menambah jumlah soal kategori mudah agar distribusi menjadi lebih seimbang. Kesesuaian dengan teori ini menjadi penting untuk memastikan bahwa instrumen tes tidak hanya memiliki variasi tingkat kesukaran, tetapi juga mampu mengukur kemampuan peserta didik secara komprehensif dan proporsional (Sudjana, 2010).

Analisis Butir Soal Ekstrem

Selain melihat distribusi tingkat kesukaran secara umum, analisis terhadap butir soal ekstrem—yaitu soal yang paling mudah dan paling sulit—menjadi langkah penting untuk memahami kualitas instrumen tes secara lebih mendalam. Butir soal ekstrem dapat memberikan informasi spesifik mengenai karakteristik soal, baik dari segi materi, konstruksi, maupun tingkat keterpahaman oleh peserta didik. Dalam penelitian ini, identifikasi butir soal ekstrem didasarkan pada nilai indeks kesukaran tertinggi dan terendah yang diperoleh dari hasil perhitungan sebelumnya (Arikunto, 2009).

Berdasarkan hasil analisis, butir soal yang termasuk dalam kategori paling mudah adalah soal nomor 1 dengan indeks kesukaran sebesar 0,72. Nilai ini menunjukkan bahwa sebagian besar mahasiswa mampu menjawab soal tersebut dengan benar. Tingginya tingkat keberhasilan ini mengindikasikan bahwa soal tersebut memiliki tingkat kesulitan yang relatif rendah. Hal ini dapat disebabkan oleh beberapa faktor, di antaranya materi yang diujikan bersifat dasar atau sudah sangat familiar bagi mahasiswa, struktur kalimat yang sederhana, serta pilihan jawaban (distraktor) yang kurang efektif dalam mengecoh peserta tes. Dalam perspektif evaluasi pembelajaran, soal yang terlalu mudah cenderung memiliki daya pembeda yang rendah, karena hampir semua peserta dapat menjawabnya dengan benar tanpa adanya perbedaan yang signifikan dalam tingkat kemampuan (Sudijono, 2011).

Meskipun demikian, keberadaan soal mudah tetap memiliki peran penting dalam suatu instrumen tes. Soal jenis ini dapat berfungsi sebagai pengantar yang membantu peserta didik membangun kepercayaan diri dalam mengerjakan tes. Selain itu, soal mudah juga dapat digunakan untuk mengukur penguasaan materi dasar yang seharusnya telah dikuasai oleh seluruh peserta didik. Oleh karena itu, meskipun soal nomor 1 tergolong sangat mudah, keberadaannya tetap relevan dalam konteks penyusunan instrumen tes yang seimbang (Sudjana, 2010). Di sisi lain, butir soal yang termasuk dalam kategori paling sulit dalam penelitian ini adalah soal nomor 14 dengan indeks kesukaran sebesar 0,16, diikuti oleh soal nomor 13 dengan indeks kesukaran sebesar 0,20. Nilai ini menunjukkan bahwa hanya sebagian kecil mahasiswa yang mampu menjawab soal tersebut dengan benar. Rendahnya tingkat keberhasilan ini mengindikasikan bahwa soal tersebut memiliki tingkat kesulitan yang tinggi. Beberapa kemungkinan penyebabnya antara lain kompleksitas materi yang diujikan, penggunaan kaidah bahasa yang lebih mendalam, atau konstruksi soal yang menuntut kemampuan analisis yang tinggi dari peserta didik (Mardapi, 2008).

Selain faktor materi, kualitas distraktor juga dapat memengaruhi tingkat kesukaran suatu soal. Distraktor yang baik adalah distraktor yang mampu mengecoh peserta didik yang belum memahami materi dengan baik, sehingga hanya peserta didik dengan pemahaman yang kuat yang dapat memilih jawaban yang benar. Dalam kasus soal nomor 13 dan 14, kemungkinan besar distraktor yang digunakan cukup efektif, sehingga banyak mahasiswa yang memilih jawaban yang salah. Hal ini menunjukkan bahwa dari segi konstruksi, soal tersebut memiliki kualitas yang cukup baik dalam hal kemampuan mengecoh peserta didik (Alderson et al., 1995). Namun demikian, soal yang terlalu sulit juga perlu dievaluasi secara kritis. Jika suatu soal memiliki tingkat kesukaran yang sangat tinggi, terdapat kemungkinan bahwa soal tersebut tidak hanya mengukur kemampuan yang diharapkan, tetapi juga dipengaruhi oleh faktor lain, seperti redaksi soal yang kurang jelas atau materi yang belum diajarkan secara optimal. Oleh karena itu, butir soal dengan tingkat kesukaran yang sangat rendah perlu ditinjau kembali untuk memastikan bahwa soal tersebut benar-benar valid dalam mengukur kemampuan yang dimaksud (Sugiyono, 2013).

Analisis terhadap butir soal ekstrem ini memberikan gambaran yang lebih rinci mengenai kualitas instrumen tes, khususnya dalam mengidentifikasi butir soal yang perlu dipertahankan, direvisi, atau bahkan diganti. Soal yang terlalu mudah dapat diperbaiki dengan meningkatkan kualitas distraktor atau menambahkan unsur kompleksitas, sedangkan soal yang terlalu sulit dapat disederhanakan atau disesuaikan dengan tingkat kemampuan peserta didik. Dengan demikian, analisis ini memiliki peran penting dalam proses penyempurnaan instrumen tes agar lebih efektif dalam mengukur kemampuan mahasiswa (Sudjana, 2010). Secara keseluruhan, keberadaan butir soal ekstrem dalam instrumen tes TOAFL aspek *tarākīb* menunjukkan adanya variasi tingkat kesukaran yang cukup jelas. Variasi ini merupakan salah satu indikator bahwa instrumen tes telah dirancang dengan mempertimbangkan perbedaan tingkat kemampuan peserta didik. Namun demikian, proporsi dan kualitas butir soal ekstrem tetap perlu diperhatikan agar tidak mengganggu keseimbangan distribusi tingkat kesukaran secara keseluruhan.

Analisis Distraktor dan Kajian Linguistik Butir Soal Representatif

Untuk memperoleh pemahaman yang lebih mendalam mengenai penyebab variasi tingkat kesukaran butir soal, analisis kuantitatif perlu dilengkapi dengan kajian terhadap distraktor serta karakteristik linguistik pada beberapa soal representatif. Dalam konteks tes tarākīb, tingkat kesukaran suatu butir tidak hanya dipengaruhi oleh banyaknya mahasiswa yang menjawab benar atau salah, tetapi

juga oleh kompleksitas kaidah nahwu-sharaf yang diujikan serta kualitas pilihan pengecoh yang disusun.

Salah satu contoh butir soal yang tergolong relatif mudah adalah soal dengan redaksi **بني الإسلام ..... خمس**dengan pilihan jawaban berupa beberapa huruf jar. Butir ini cenderung mudah karena mengacu pada ungkapan hadis yang sangat familiar, yaitu *بني الإسلام على خمس*. Selain faktor familiaritas teks, mahasiswa juga cukup terbantu dengan adanya petunjuk kaidah huruf jar yang secara langsung menuntun pada jawaban yang benar. Distraktor pada soal ini kurang memiliki kekuatan pengecoh karena pilihan selain **على**secara struktur tidak lazim digunakan dalam susunan tersebut. Kondisi ini menjelaskan mengapa soal dengan karakter seperti ini lebih mudah dijawab oleh mayoritas mahasiswa.

Sebaliknya, beberapa butir soal yang tergolong sulit menunjukkan bahwa faktor kesulitan tidak semata-mata berasal dari materi, tetapi juga dari efektivitas distraktor yang sangat dekat secara morfologis maupun sintaktis. Contohnya adalah soal **يسرني مساعدتك .....** dengan pilihan **أخاك، أخيك، أخوك، أخك**. Soal ini menuntut mahasiswa memahami posisi kata sebagai *maf'ul bih* dalam bentuk *isim khamsah* yang dimudhafkan, sehingga jawaban yang tepat adalah **أخاك**. Kesulitan muncul karena seluruh pilihan tampak serupa secara leksikal, namun berbeda pada tanda i'rab. Distraktor seperti ini sangat efektif mengecoh mahasiswa yang belum menguasai perubahan akhir kata berdasarkan fungsi sintaksis.

Contoh lain terdapat pada butir soal berbentuk identifikasi kesalahan kaidah, seperti kalimat **إن التي الطالبة يجتهد في دروسها ستنجح في الامتحان النهائي**. Kesulitan soal ini terletak pada tuntutan analisis lebih dari satu unsur sekaligus, yakni ketidaksesuaian antara *ism maushul* muannats, subjek perempuan, serta penggunaan fi'il mudhari' yang seharusnya menyesuaikan bentuk muannats menjadi *تجتهد*. Distraktor dalam soal model ini tidak berbentuk pilihan jawaban sederhana, tetapi berupa beberapa kata bergaris bawah yang semuanya tampak benar secara makna, sehingga mahasiswa harus benar-benar jeli pada aspek kaidah.

Demikian pula pada kalimat **قرأ عليا القران الكريم بعد كل صلاة**yang menguji ketepatan i'rab isim 'alam. Banyak mahasiswa mengalami kesulitan karena harus membedakan antara bentuk nominatif dan akusatif pada nama *علي*. Bentuk **عليا**pada kalimat tersebut merupakan konstruksi yang tidak tepat, namun secara visual sering dianggap benar oleh mahasiswa yang belum kuat dalam penguasaan tanda i'rab.

Berdasarkan contoh-contoh tersebut dapat dipahami bahwa butir soal yang tergolong sulit pada instrumen TOAFL aspek tarākīb umumnya memiliki dua karakteristik utama, yaitu: (1) distraktor yang sangat mirip dengan jawaban benar sehingga menuntut ketelitian tinggi, dan (2) keterlibatan lebih dari satu kaidah nahwu-sharaf dalam satu soal. Dengan demikian, tingkat kesukaran dalam tes ini tidak hanya merefleksikan rendahnya penguasaan mahasiswa, tetapi juga menunjukkan bahwa penyusun soal telah menggunakan pengecoh linguistik yang cukup efektif.

Analisis Daya Pembeda

Berdasarkan hasil perhitungan daya pembeda, diperoleh bahwa kualitas diskriminatif butir soal menunjukkan variasi yang cukup signifikan. Dari 20 butir soal yang dianalisis, terdapat 10 butir soal dengan kategori sangat baik, 4 butir soal berkategori cukup, dan 6 butir soal berkategori kurang. Temuan ini menunjukkan bahwa sebagian butir soal telah memiliki kemampuan yang optimal dalam membedakan mahasiswa berkemampuan tinggi dan rendah, meskipun masih terdapat beberapa soal yang belum efektif.

Jika dikaitkan dengan tingkat kesukaran, butir-butir soal yang berada pada kategori sedang cenderung memiliki daya pembeda lebih baik dibandingkan dengan soal yang terlalu mudah maupun terlalu sulit. Hal ini menunjukkan bahwa tingkat kesukaran yang moderat memberikan peluang lebih besar bagi soal untuk bekerja secara diskriminatif. Sementara itu, beberapa soal yang memiliki daya pembeda rendah perlu direvisi agar mampu menjalankan fungsi evaluatif secara lebih maksimal

Analisis Reliabilitas Tes

Hasil perhitungan reliabilitas menggunakan rumus KR-20 menunjukkan bahwa koefisien reliabilitas tes sebesar 0,55. Nilai ini berada pada kategori reliabilitas sedang, yang menandakan bahwa instrumen tes memiliki konsistensi internal yang cukup dalam mengukur kemampuan mahasiswa pada aspek tarākīb. Meskipun demikian, nilai tersebut menunjukkan bahwa instrumen masih memerlukan penyempurnaan agar tingkat keandalannya meningkat.

Reliabilitas yang belum optimal ini dipengaruhi oleh masih adanya beberapa butir soal dengan daya pembeda rendah dan distribusi tingkat kesukaran yang belum proporsional. Dengan demikian, perbaikan pada butir-butir soal yang kurang efektif berpotensi meningkatkan kualitas instrumen secara keseluruhan.

Implikasi terhadap Kualitas Instrumen dan Rekomendasi

Berdasarkan hasil analisis tingkat kesukaran yang telah dilakukan, dapat disimpulkan bahwa instrumen tes TOAFL pada aspek *tarākīb* dalam penelitian ini secara umum telah memiliki kualitas yang cukup baik, terutama dari segi variasi tingkat kesukaran. Hal ini ditunjukkan dengan adanya butir soal yang mencakup kategori mudah, sedang, dan sulit. Variasi ini penting dalam suatu instrumen evaluasi, karena memungkinkan pengukuran kemampuan peserta didik secara lebih komprehensif, mulai dari kemampuan dasar hingga kemampuan tingkat tinggi. Dengan demikian, instrumen tes ini dapat dikatakan telah memenuhi salah satu prinsip dasar dalam penyusunan alat evaluasi yang baik.

Namun demikian, jika ditinjau dari segi keseimbangan distribusi tingkat kesukaran, instrumen tes ini masih belum sepenuhnya memenuhi kriteria ideal. Proporsi soal yang didominasi oleh kategori sedang (70%) dan cukup banyaknya soal sulit (25%) menunjukkan bahwa instrumen tes cenderung lebih menekankan pada pengukuran kemampuan menengah hingga tinggi. Sementara itu, jumlah soal mudah yang sangat terbatas (5%) menunjukkan adanya kekurangan dalam mengakomodasi peserta didik dengan kemampuan rendah. Kondisi ini dapat berdampak pada kurang optimalnya fungsi tes dalam memberikan gambaran menyeluruh mengenai kemampuan mahasiswa.

Implikasi dari temuan ini menunjukkan bahwa meskipun instrumen tes telah mampu mengukur kemampuan mahasiswa secara cukup baik, namun masih diperlukan penyempurnaan agar hasil evaluasi menjadi lebih akurat dan proporsional. Instrumen tes yang tidak seimbang dalam distribusi tingkat kesukarannya berpotensi menghasilkan data yang bias, terutama dalam mengukur kemampuan peserta didik pada tingkat tertentu. Oleh karena itu, perbaikan dalam penyusunan soal menjadi hal yang penting untuk dilakukan. Temuan dalam penelitian ini juga menunjukkan bahwa analisis instrumen tes tidak dapat hanya bergantung pada satu indikator, seperti tingkat kesukaran, tetapi perlu mempertimbangkan indikator lain seperti daya pembeda dan reliabilitas untuk memperoleh gambaran kualitas yang lebih komprehensif. Hal ini memperkuat pandangan dalam teori evaluasi pendidikan bahwa kualitas instrumen ditentukan oleh kombinasi beberapa parameter psikometrik, bukan hanya satu aspek saja.

Kontribusi utama penelitian ini terletak pada penyajian model analisis sederhana namun komprehensif yang dapat diterapkan oleh dosen atau penyusun soal dalam mengevaluasi instrumen tes TOAFL, khususnya pada aspek tarākīb. Model ini dapat menjadi acuan praktis dalam melakukan evaluasi berkala terhadap kualitas soal, sehingga proses penyusunan instrumen tidak bersifat statis, melainkan berbasis data empiris. Dengan demikian, penelitian ini tidak hanya memberikan deskripsi mengenai tingkat kesukaran butir soal, tetapi juga menawarkan pendekatan evaluatif yang lebih sistematis dalam pengembangan instrumen tes bahasa Arab. Kontribusi ini diharapkan dapat menjembatani kebutuhan antara kajian teoretis dalam evaluasi pembelajaran dan praktik penyusunan soal di lapangan.

Adapun beberapa rekomendasi yang dapat diajukan berdasarkan hasil penelitian ini adalah sebagai berikut. Pertama, perlu dilakukan penambahan jumlah butir soal kategori mudah agar distribusi tingkat kesukaran menjadi lebih seimbang. Soal mudah berperan penting dalam mengukur kemampuan dasar serta memberikan kesempatan bagi seluruh peserta didik untuk menunjukkan kemampuannya. Kedua, soal kategori sedang yang sudah mendominasi perlu dipertahankan, karena jenis soal ini memiliki kemampuan terbaik dalam mengukur variasi kemampuan peserta didik. Ketiga, soal kategori sulit tetap perlu dipertahankan dalam jumlah yang proporsional, namun perlu dilakukan evaluasi terhadap butir soal yang terlalu sulit untuk memastikan bahwa soal tersebut benar-benar mengukur kemampuan yang diharapkan dan bukan disebabkan oleh faktor lain seperti redaksi yang kurang jelas. Selain itu, penyusun soal disarankan untuk melakukan analisis butir soal secara berkala, tidak hanya terbatas pada tingkat kesukaran, tetapi juga mencakup aspek lain seperti daya pembeda dan validitas butir soal. Analisis yang komprehensif akan membantu dalam meningkatkan kualitas instrumen tes secara keseluruhan, sehingga evaluasi pembelajaran dapat dilakukan secara lebih efektif dan akurat.

Dengan demikian, instrumen tes yang digunakan tidak hanya berfungsi sebagai alat ukur hasil belajar, tetapi juga sebagai alat refleksi dalam meningkatkan kualitas pembelajaran.

Dengan mempertimbangkan berbagai implikasi dan rekomendasi tersebut, penelitian ini diharapkan dapat memberikan kontribusi praktis bagi pengembangan instrumen evaluasi pembelajaran bahasa Arab, khususnya dalam konteks tes TOAFL. Upaya perbaikan yang berkelanjutan terhadap instrumen tes akan berdampak positif terhadap kualitas evaluasi pembelajaran, yang pada akhirnya juga akan meningkatkan kualitas proses dan hasil pembelajaran itu sendiri.

Keterbatasan Penelitian

Penelitian ini memiliki beberapa keterbatasan yang perlu dicermati dalam menafsirkan hasil temuan. Pertama, jumlah subjek penelitian yang hanya melibatkan 25 mahasiswa masih tergolong terbatas untuk memberikan generalisasi yang kuat terhadap keseluruhan kualitas instrumen TOAFL sebagai tes standar bahasa Arab. Dengan jumlah sampel yang relatif kecil, hasil penelitian ini lebih tepat dipahami sebagai gambaran empiris pada konteks lokal penggunaan instrumen daripada sebagai representasi menyeluruh terhadap karakteristik tes TOAFL.

Kedua, data penelitian hanya diambil dari satu periode pelaksanaan tes, yaitu tahun 2021, sehingga belum dapat menunjukkan konsistensi kualitas instrumen apabila digunakan pada periode yang berbeda atau pada populasi mahasiswa yang lebih luas. Ketiga, penelitian ini berfokus pada analisis kuantitatif butir soal dan belum menyertakan analisis kualitatif terhadap redaksi soal, efektivitas distraktor, serta kesesuaian materi dengan indikator kompetensi tarākīb.

Oleh karena itu, penelitian lanjutan disarankan untuk melibatkan jumlah responden yang lebih besar, data lintas periode, serta analisis instrumen yang lebih mendalam agar diperoleh gambaran kualitas tes TOAFL yang lebih komprehensif.

KESIMPULAN

Berdasarkan hasil analisis terhadap 20 butir soal TOAFL aspek tarākīb yang diujikan kepada 25 mahasiswa, dapat disimpulkan bahwa kualitas instrumen menunjukkan variasi yang cukup beragam. Ditinjau dari tingkat kesukaran, butir soal tersebar ke dalam kategori mudah, sedang, dan sulit dengan dominasi pada kategori sedang, sehingga secara umum instrumen telah memiliki distribusi kesulitan yang cukup proporsional untuk mengukur kemampuan struktur bahasa Arab mahasiswa.

Dari aspek daya pembeda, diperoleh bahwa 10 butir soal berkategori sangat baik, 4 butir soal berkategori cukup, dan 6 butir soal berkategori kurang. Hasil ini menunjukkan bahwa sebagian besar butir soal telah memiliki kemampuan diskriminatif yang memadai dalam membedakan mahasiswa berkemampuan tinggi dan rendah, meskipun beberapa soal masih memerlukan revisi. Sementara itu, hasil perhitungan reliabilitas menggunakan KR-20 menunjukkan koefisien sebesar 0,55 yang berada pada kategori sedang, sehingga instrumen dapat dinilai cukup konsisten sebagai alat evaluasi.

Analisis distraktor dan kajian linguistik terhadap beberapa butir soal representatif menunjukkan bahwa tingkat kesukaran soal dipengaruhi oleh efektivitas pengecoh, kemiripan bentuk morfologis antar pilihan jawaban, serta keterlibatan lebih dari satu kaidah nahwu-sharaf dalam satu soal. Dengan demikian, kualitas instrumen TOAFL aspek tarākīb dapat dinilai cukup memadai, namun tetap memerlukan penyempurnaan pada beberapa butir soal agar fungsi pengukuran kemampuan bahasa Arab menjadi lebih optimal.

DAFTAR PUSTAKA

Alderson, Charles, J., Clapham, C., & Wall, D. (1995). *Language test construction and evaluation*.
Anastasi, A., & Urbina, S. (1997). *Psychological testing* (7th ed.). Prentice Hall.
Arikunto, S. (2009). *DASAR - DASAR EVALUASI PENDIDIKAN* (p. 310). https://www.scribd.com/document/740858744/Dasar-dasar-Evaluasi-Pendidikan
Azwar, S. (2012). *Reliabilitas dan validitas*. Pustaka Pelajar.
Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice*. Oxford University Press.
Bachman, L. F. (2004). *Statistical analyses for language assessment*. Cambridge University Press.

Brookhart, S. M. (2013). *How to create and use rubrics for formative assessment and grading*. ASCD.
Brown, H. D. (2004). *Language assessment: Principles and classroom practices*. Pearson Education.
Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. Holt, Rinehart & Winston.
Fulcher, G., & Davidson, F. (2007). *Language testing and assessment: An advanced resource book*. Routledge
Gronlund, N. E., & Linn, R. L. (1990). *Measurement and evaluation in teaching* (6th ed.). Macmillan.
Haladyna, T. M. (2004). *Developing and validating multiple-choice test items* (3rd ed.). Lawrence Erlbaum Associates.
Mardapi, D. (2008). *Teknik penyusunan instrumen tes dan nontes*.
McNamara, T. (2000). *Language testing*. Oxford University Press.
Nitko, A. J., & Brookhart, S. M. (2011). *Educational assessment of students* (6th ed.). Pearson.
Oller, J. W. (1979). *Language tests at school*. Longman.
Purpura, J. E. (2004). *Assessing grammar*. Cambridge University Press.
Retnawati, H. (2016). *Analisis kuantitatif instrumen penelitian*. Parama Publishing.
Sudijono, A. (2011). *PENGANTAR EVALUASI PENDIDIKAN* (p. 480).
Sudjana, N. (2010). *Penilaian hasil proses belajar mengajar*.
Sugiyono. (2013). *Metode Penelitian Pendidikan pendekatan kuantitatif, kualitatid, dan R&D*. ALFABETA, CV.
Sukmadinata, & Syaodih, N. (2002). *Pengembangan kurikulum teori dan praktek*.