

## KUALITAS BUTIR SOAL UAS BAHASA INGGRIS UNTUK SISWA MTs DI SAMARINDA

**Sumaningsih**  
**MTs Almuna Samarinda**  
**Email: [suma.almuna@gmail.com](mailto:suma.almuna@gmail.com)**

**Abstract:** This study is aimed at seeing the item quality of English test used for Leaving Exam at SMP Samarinda. Analysis is used to see item difficulty (p), discrimination (D), distracter effectiveness, validity, and reliability. The results of analysis show that of 45 items, 19 (42.2%) must be dropped due to the poor p and 17 (37.8%) receive the low D. Analysis of p and D reveal that of 45 items, only 5 (11.1%) items are applicable, 21 (46.7%) need revision, and 19 (42.2%) should be dropped. The distracters are also poor, as 25 (55.6%) items have very poor distracters. Validity testing under Product Moment formula achieved low value (0.346) and reliability testing under KR-21 shows fair value (0.650).

**Keywords:** UAS, English test, quality.

**TINGKAT** pencapaian hasil UAS di Kaltim diklaim para pendidik dan orang tua murid kurang berhasil. Setiap tahun berdasarkan hasil penerimaan Danem terendah di SMP dan SMA, dianggap mengalami penurunan. Tetapi, SMP dan SMA Negeri selalu kelebihan pendaftar dengan NEM di atas 33.

Persoalan ini secara sepintas dianggap sebagai akibat menurunnya kemampuan siswa. Tanpa didukung dengan bukti-bukti yang akurat mengenai keadaan lingkungan dan faktor lain yang mungkin mempengaruhi, siswa dipandang kurang mampu mengerjakan UAS. Indikator utama kekurang-berhasilan itu ialah rata-rata NEM setiap tahun berada di bawah 6.0. Bahkan untuk sekolah-sekolah swasta mencapai 3.0 hingga 4.0.

Penelitian mengenai hasil UAS di Samarinda sejauh ini baru dilakukan terhadap produk akhirnya berupa NEM. Kecurigaan terhadap proses penulisan butir soal baru dilakukan oleh Budiharso (1994) mengenai validitas dan reliabilitas tes bahasa Inggris SMA, Riandy (1997) mengenai analisis soal bahasa Inggris UAS SMP, dan Budiharso (1997) mengenai kualitas item tes UAS bahasa Inggris. Ketiga penelitian tersebut menunjukkan bahwa baik butir-butir setiap soal maupun butir soal secara keseluruhan tes bahasa Inggris UAS memang rendah.

Penelitian ini dimaksudkan untuk mengetahui kualitas butir soal UAS Bahasa Inggris untuk SMP di Samarinda tahun 2014/2015 dalam hal (1) tingkat kesulitan, (2) daya pembeda, (3) efektivitas distraktor, (4) validitas, dan (5) reliabilitas.

## **TES BUATAN GURU DAN TES TERSTANDAR**

Tes buatan guru menurut Harris (1969:1) ialah tes yang disiapkan, dilaksanakan, dan diskor oleh seorang guru untuk mengukur pencapaian siswanya sendiri. Tujuan pengetesan ialah tujuan pengajaran, sedangkan materi tes diambil dari bahan pengajaran juga. Atas dasar kutipan di atas, bisa disimpulkan bahwa tes buatan guru memiliki ketepatan materi tes dan tujuan apa yang sudah diberikan di kelas. Tes buatan guru digunakan untuk mengukur sejauh mana kemajuan siswa dan mendiagnosa kekuatan dan kelemahan siswa dalam penguasaan materi tertentu.

Pada umumnya, pengertian tes buatan guru selalu dihubungkan dengan tes terstandar. Tes terstandar menurut Ebel (1986:267) ialah tes yang (1) tes yang disusun oleh para ahli penyusun tes dan telah diujicobakan, dianalisis, dan diperbaiki, (2) memiliki petunjuk pelaksanaan dan pemberian skor yang jelas, dan (3) memiliki acuan norma untuk penginterpretasian skor. Tes terstandar dapat dipakai berulang-ulang di berbagai sekolah bahkan di berbagai negara.

Dengan membandingkan dua definisi di atas, bisa disimpulkan bahwa tes yang membedakan secara tegas antara tes buatan guru dengan tes terstandar terutama pada proses ujicoba dan analisis butir soalnya. Tes buatan guru karena digunakan untuk kalangan sendiri, sering tidak memerlukan ujicoba. Sebaliknya, tes terstandar memerlukan ujicoba sebagai kriteria utama untuk menyeleksi butir soal yang baik.

Karena tidak diujicobakan, tes buatan guru tidak dilengkapi informasi mengenai kualitas butir, validitas dan reliabilitas soal. Begitu juga sistem skoringnya, sering tergantung pada pola yang diterapkan guru sendiri. Dengan demikian, walaupun para guru bergabung untuk menyusun soal, tetapi apabila proses ujicoba, revisi soal, dan analisis butir belum dilaksanakan, tes tersebut masih terkatagori pada tes buatan guru.

## **CIRI-CIRI TES YANG BAIK**

Menurut Harris (1969:13) dan Djiwandono (1996:90) sebuah tes dikatakan baik apabila tes tersebut memenuhi persyaratan utama yaitu: valid, reliabel, dan praktis. Tes yang baik merupakan tuntutan pokok dalam mengetahui hasil belajar bahasa. Menurut Djiwandono (1996:90) melalui penyelenggaraan tes, diharapkan dapat diperoleh informasi mengenai berbagai segi penyelenggaraan pengajaran bahasa. Di samping hasil informasi mengenai hasil belajar siswa sebagai salah satu segi yang terpenting, penyelenggaraan tes dalam pengajaran bahasa memberikan pula informasi tentang ketepatan identifikasi tujuan, kesesuaian bahan pengajaran, kesesuaian dan keefektifan metode pengajaran, cukup tidaknya latihan yang diberikan, kesulitan belajar siswa. Semua informasi itu dapat diperoleh atas dasar pengamatan terhadap tingkat keberhasilan siswa seperti tercermin pada nilai-nilai yang dicapai, dan kajian terhadap tingkat dan jenis kesalahan yang dibuat oleh peserta tes. Mengingat begitu pentingnya fungsi tes, maka betapa perlunya mengusahakan agar tes yang digunakan dalam pengajaran benar-benar merupakan alat yang baik dan dapat diandalkan.

## **VALIDITAS**

Validitas suatu tes diartikan sebagai kesesuaian antara skor yang diperoleh siswa dengan tujuan yang hendak diuji (Gronlund, 1979; Ebel, 1979; Djiwandono, 1996). Dengan

demikian, yang dimaksud dengan valid bukanlah tesnya, tetapi skor yang diperoleh siswa sudah sesuai dengan tujuan tes atau belum. Andaikan soal yang ditulis dimaksudkan untuk menguji kemampuan membaca, maka butir-butir yang ditulis untuk soal harus menanyakan pemahaman isi wacana. Jika ada pertanyaan yang mengharuskan penjawab soal mengenai menulis, maka butir soal tersebut tidak akan menunjukkan hasil pemahaman. Oleh sebab itu, skor yang diperoleh siswa tidak menunjukkan kemampuan yang sebenarnya mengenai membaca.

Secara umum, validitas dikategorikan ke dalam dua jenis: validitas logis dan validitas empiris (Gronlund, 1979). Validitas logis, menurut Djiwandono (1996) diperoleh apabila unsur-unsur pembuktiannya memerlukan penalaran. Sebaliknya, validitas empiris memerlukan pembuktian melalui tes statistik dan memerlukan adanya kriteria sebagai pembanding.

Jenis-jenis validitas, karena itu bisa dibedakan menjadi empat jenis: validitas tampak luar, validitas isi, validitas kriteria, validitas konstruk dan validitas (Djiwandono, 1996; Gronlund, 1979; Oller, 1979). Sejauh mana sebuah tes memiliki tingkat validitas yang tinggi, diukur dengan seberapa banyak evidence yang bisa digunakan untuk mendukung kualitas tes. Misalnya, sebuah tes yang memiliki evidence isi, kriteria dan konstruk, akan lebih meyakinkan dibanding tes yang hanya memiliki evidence dalam hal validitas kriteria saja.

Proses kemantapan evidence untuk mendukung validitas dimulai ketika butir-butir soal dirancang dalam bentuk kisi-kisi (*Table of Specification*) yang berisi proporsi tingkat kesulitan dan penyebaran materi yang akan diuji, penulisan butir soal dan hasil try-out untuk mengetahui tingkat kesulitan dan daya beda setiap butir soal. Apabila soal tersebut berbentuk pilihan ganda, dua hal tersebut masih harus ditambah dengan efektivitas option (distractor).

Apabila validasi suatu tes dilakukan menggunakan uji statistik, tingkat koefisien validitas berkisar antara 0-1. Semakin mendekati angka 1, hasil tes semakin valid. Ebel (1979) memberikan angka .50 untuk tes yang digunakan di kelas dan .90 ke atas untuk tes terstandar.

Harris (1969:19) menyebutkan bahwa tes dikatakan valid apabila tes tersebut melakukan dua hal secara baik: (1) mengukur secara tepat apa yang hendak diukur dan tidak mengukur hal lain yang bukan menjadi tujuannya, dan (2) seberapa baik tes tersebut mengukur tujuan tersebut. Menurut Djiwandono (1996:91) validitas merupakan ciri tes yang amat penting, yang seharusnya dimiliki oleh setiap tes yang digunakan untuk berbagai tujuan, termasuk tujuan pengajaran bahasa. Secara konvensional, validitas diartikan sebagai ciri yang menunjukkan adanya kesesuaian antara tes dengan apa yang ingin diukur dengan menggunakan tes itu. Secara lebih tepat, ditegaskan oleh Djiwandono, bahwa sebenarnya validitas bukan merupakan ciri yang terkait pada tesnya sebagai alat, tetapi apakah nilai yang diperoleh peserta tes bisa menunjukkan kemampuan mengenai bahan yang diujikan.

Dengan demikian, validitas tes tergantung pada beberapa hal. Pertama, validitas tergantung pada tujuan pengadaan tes. Untuk apa tes itu dilakukan; untuk menempatkan siswa, menjarang siswa, melihat kemampuan siswa. Kedua, validitas tes tergantung pada kelompok yang dites. Suatu tes yang valid untuk anak-anak, akan menjadi tidak valid untuk orang dewasa. Ketiga, validitas tes tergantung juga pada situasi dan kondisi saat tes itu digunakan.

Dewi (1991:26-30) membedakan validitas ke dalam dua aspek: aspek kuantitatif dan aspek kualitatif. Secara garis besar, aspek kuantitatif digolongkan ke dalam validitas empiris (empirical validity), sedangkan aspek kualitatif dikelompokkan ke dalam validitas logis (logical validity).

### **Validitas Empiris**

Validitas empiris ialah validitas suatu tes yang dicapai dengan cara membandingkan atau mengkorelasikan skor tes dengan skor tes lain yang dipakai sebagai kriteria. Yang termasuk dalam validitas empiris ini ialah (1) validitas peramalan (predictive validity), dan (2) validitas bandingan (concurrent validity). Validitas peramalan menunjukkan apakah tes yang dikerjakan sekarang dapat meramalkan performansi seseorang di masa yang akan datang (Tuckman, 1975:238).

Misalnya, suatu sekolah mengadakan dua jenis tes, yaitu tes bakat membaca yang diteskan pada semester 2, dan tes membaca pemahaman yang diteskan pada semester ke 4. Setelah skor kedua tes dikorelasikan, diketahui bahwa hasilnya 0.58. Angka ini menunjukkan bahwa tes bakat mempunyai validitas ramalan untuk membaca pemahaman karena tes tersebut bisa meramalkan kemampuan siswa di bidang membaca.

Satu lagi validitas empiris ialah validitas bandingan. Validitas bandingan ialah validitas tes ditinjau dari korelasinya dengan dengan kemampuan yang dimiliki saat ini. Kemampuan (skor kemampuan) yang ada pada saat tes dilaksanakan dapat diperoleh dari nilai dalam rapor semester saat itu, dari hasil tes lain yang sejenis yang memiliki validitas, atau dari guru yang membina siswa tersebut.

Dengan demikian, validitas logis pembuktiannya memerlukan alat berupa tes. Dua metode yang sering digunakan untuk menguji validitas empiris ini di antaranya (1) metode rank-difference untuk sejumlah skor kecil; dan (2) metode product moment untuk sejumlah skor besar. Nilai validitas ini dinyatakan dengan koefisien dan besarnya berkisar antara 0 hingga 1 (Djiwandono, 1996:95). Koefisien yang nilainya negatif atau lebih dari 1 menunjukkan penyimpangan antara tujuan yang akan dites dan skor yang ditunjukkan peserta tes (Ebel, 1979).

### **Validitas Logis**

Validitas logis merupakan pembuktian kesesuaian isi pertanyaan dengan apa yang hendak diukur berdasarkan pengertian-pengertian logika. Karena itu, validitas ini tidak memerlukan alat uji berupa tes untuk menentukan kriterianya. Yang termasuk validitas logis ialah (1) validitas isi (content validity), (2) validitas konstruk (construct validity), dan (3) validitas penampakan (face validity). Yang dimaksud validitas isi ialah validitas suatu bentuk tes dilihat dari isi atau materi tes tersebut. Suatu tes memiliki validitas isi apabila materi/isi tes tersebut merupakan bahan-bahan yang mewakili materi yang sudah diajarkan. Validitas isi bisa dianalisis menggunakan kisi-kisi soal.

Sebuah tes dikatakan memiliki validitas konstruk, menurut Heaton (1984:154) apabila tes tersebut dapat diukur ciri-ciri khusus tertentu yang sesuai dengan teori tingkah laku bahasa dan teori belajar bahasa. Misalnya ada asumsi yang mengatakan bahwa kebiasaan berbahasa

dapat diperoleh dari belajar grammar, maka tes yang memiliki validitas konstruk hendaknya disusun atas dasar penguasaan grammar.

Yang dimaksud validitas penampakan ialah ciri tes yang menunjukkan apakah suatu tes nampaknya mengukur apa yang hendak diukur (Djiwandono, 1996:97). Dalam penerapannya, ciri itu berupa hal-hal yang secara tampak luar memberikan kesan bahwa suatu tes memang dapat melakukan pengukuran yang seharusnya diukur. Misalnya, tes berbicara dikatakan memiliki validitas penampakan apabila peserta tes nampak saling berbicara dan bukan menulis.

## RELIABILITAS

Reliabilitas diartikan sebagai keajegan atau kekonsistensian skor tes apabila tes tersebut diberikan kepada peserta lain yang memiliki tingkat kemampuan yang (hampir) sama (Gronlund, 1979; Oller, 1979; Weir, 1983; Djiwandono, 1996). Sebuah tes dikatakan tidak reliabel apabila setelah diberikan kepada peserta tes lain yang memiliki kemampuan hampir sama pada kesempatan yang berbeda, menghasilkan skor yang jauh berbeda (Brown, 1987).

Berbeda dengan validitas yang bisa diuji melalui analisis isi, reliabilitas menuntut piranti statistik untuk membuktikannya. Teknik pengujian reliabilitas ini mencakup dua hal: reliabilitas untuk soal yang memiliki skor dikotomus dan soal yang diskor secara berskala. Soal-soal yang dikotomus (salah = 0, benar = 1) diuji reliabilitasnya menggunakan metode tes ulang, metode tes setara, metode belah dua, metode Kuder-Richardson. Sedangkan soal essay yang diskor secara berskala diuji reliabilitasnya menggunakan metode koefisien alfa, dan metode antarpemilai (Djiwandono, 1996; Oller, 1979; Gronlund, 1979).

Tingkat koefisien reliabilitas juga bergerak dari 0-1. Semakin mendekati angka 1 semakin reliabel. Namun reliabilitas 1 dalam praktik hampir mustahil terjadi (Gronlund, 1979; Ebel, 1979; Djiwandono, 1996). Koefisien reliabilitas dianggap sudah sangat baik apabila mencapai angka .90 ke atas. Jika terdapat koefisien negatif atau lebih dari satu, baik untuk validitas maupun reliabilitas, maka pasti terjadi proses yang salah.

Dengan kata lain, reliabilitas diartikan sebagai ciri tes yang memiliki kemampuan untuk menghasilkan pengukuran yang ajeg, tidak berubah-ubah seandainya digunakan secara berulang-ulang pada sasaran yang sama (Djiwandono, 1996:98). Karena itu, reliabilitas bukan terkait dengan tesnya melainkan terkait dengan skor yang ajeg yang diperoleh peserta tes. Artinya, skor yang diperoleh siswa dari tes tersebut stabil atau konsisten meskipun suasana pengujian atau penskoran yang dilakukan berbeda. Brown (1980:23) menyebutkan bahwa tes dikatakan tidak reliabel jika tes tersebut diberikan kepada peserta tes yang sama dalam waktu yang hampir sama, menghasilkan nilai yang jauh berbeda.

Berbeda dengan validitas yang pembuktiannya bisa dilakukan menggunakan logika, reliabilitas menuntut pembuktian menggunakan alat-alat statistik, misalnya menggunakan koefisien korelasi atau simpangan baku pengukuran. Djiwandono (1996:99) menyebutkan bahwa nilai koefisien reliabilitas bergerak dari 0 hingga 1. Semakin dekat dengan angka 1, semakin reliabel. Sedangkan koefisien dalam bentuk negatif atau lebih dari 1 menunjukkan reliabilitas yang menyimpang. Dalam praktik, angka reliabilitas tidak pernah mencapai angka 1. Untuk tes objektif pengujian reliabilitas di antaranya bisa dilakukan menggunakan: (1) metode tes ulang, (2) metode tes kembar, (3) metode belah dua, (4) rumus Kuder-Richardson

(KR). Sedangkan untuk tes essay metode yang digunakan ialah (1) koefisien alfa, (2) koefisien antar penilai, dan (3) tabel khusus.

Metode tes ulang mempersyaratkan penggunaan tes yang sama dua kali, pada peserta tes yang sama. Metode tes setara memerlukan dua buah tes yang setara, yang mampu menghasilkan skor yang sama, tanpa mengulang penggunaan tes yang sama. Metode belah dua diperoleh dengan cara memberi tes sekali kepada peserta tes, kemudian jawaban peserta tes dibagi menjadi dua bagian, misalnya nomor ganjil dan nomor genap. Untuk jenis reliabilitas ini, digunakan metode Spearman-Brown. Untuk tes yang diselenggarakan sekali saja, bisa dianalisis reliabilitasnya menggunakan metode KR, khususnya KR-20 atau KR-21. Penggunaan metode ini mempersyaratkan penggunaan skor diskrit, yaitu skor 1 untuk soal benar, dan skor 0 untuk soal salah.

Soal-soal essay bisa dicari reliabilitasnya menggunakan metode koefisien alfa, metode antar penilai, dan metode perkiraan. Soal yang diskor dengan cara skala memberikan nilai yang tidak sama untuk setiap butir, tergantung pada kesempurnaan jawaban dan kelengkapannya. Untuk itu, metode koefisien alfa merupakan alat analisis yang cocok. Selain itu, untuk menghindari subjektivitas bisa juga digunakan metode antar penilai. Nilai yang diberikan oleh dua orang penilai dikorelasikan.

Apabila cara statistik seperti tersebut di atas tidak bisa dilakukan, terutama untuk soal-soal buatan guru, metode analisis reliabilitas bisa dilakukan dengan cara perkiraan. Untuk ini, soal-soal dikelompokkan dalam katagori soal sulit, sedang, dan mudah. Katagori tersebut kemudian dibuat dalam tabel dan jumlah soal diproporsi sesuai dengan rencana penyusun soal.

Interpretasi koefisien validitas dan reliabilitas secara konvensional bisa dilakukan mengikuti tabel yang dikemukakan Djiandono (1996:154) berikut ini:

0.90 - 1,00	=	amat tinggi
0.70 - 0.89	=	tinggi
0.50 - 0.69	=	sedang
0.30 - 0.49	=	rendah
> 0.30	=	amat rendah

### **ANALISIS HASIL TES**

Analisis hasil tes dimaksudkan untuk mengetahui informasi mengenai kemampuan siswa tertentu dalam menjawab tes. Informasi ini berupa skor yang sudah diinterpretasi menggunakan kriteria tertentu. Untuk hasil yang memuaskan, analisis hendaknya disertai juga dengan informasi mengenai tingkat kesulitan (p) setiap butir soal, daya pembeda (D), validitas, dan reliabilitas. Untuk soal yang berbentuk pilihan ganda, analisis masih bisa ditambahkan dalam bentuk analisis efektifitas distraktor (pilihan jawaban).

### Tingkat Kesulitan (p)

Tingkat kesulitan tes menunjukkan seberapa sulit atau seberapa mudah butir-butir tes atau tes secara keseluruhan diselenggarakan (Djiwandono, 1996:140). Dengan demikian, akan diketahui apakah suatu butir tes tergolong sangat sulit, sulit, sedang, mudah, atau sangat mudah. Secara umum, tingkat kesulitan bisa dilihat dari skor rata-rata yang diperoleh siswa. Tingkat kesulitan bisa dinyatakan dalam bentuk persentase. Misalnya, tes yang dijawab benar oleh semua peserta tes, dinyatakan memiliki  $p = 100\%$  atau 1,00. Cara ini diperoleh dengan menggunakan rumus sederhana berikut:

$$p = \text{JJB} : \text{JPT} \times 100\%$$

dengan  $p$  = tingkat kesulitan, JJB = jumlah jawaban benar, sedangkan JPT = jumlah peserta tes.

Butir soal yang memiliki  $p$  mendekati 1 atau 100% berarti mudah atau sangat mudah. Sebaliknya, butir soal yang nilai  $p$ -nya mendekati 0 berarti sangat sulit atau sulit. Butir soal yang dianggap memiliki  $p$  ideal ialah 0.50, sedangkan  $p$  yang dianggap bisa diterima berkisar antara 0.20 - 0.80 (Djiwandono, 1996:141).

### Daya Pembeda (D)

Lebih lanjut, (Djiwandono, 1996:143-144) menjelaskan bahwa daya pembeda atau tingkat diskriminasi (disingkat D) menunjukkan kemampuan tes dalam membedakan kemampuan antara peserta tes yang berkemampuan tinggi dan yang berkemampuan rendah. Semakin tinggi D, semakin tinggi pula kemampuannya untuk membedakan kelompok siswa yang pandai dan yang kurang pandai. Tingkat diskriminasi 0 artinya soal tersebut tidak bisa membedakan siswa yang pandai dan yang kurang pandai. Tingkat diskriminasi  $D = 1,00$  berarti bahwa semua anggota kelompok pandai bisa menjawab benar, sementara anggota kelompok bawah tidak ada yang bisa menjawab soal dengan benar. Tingkat diskriminasi dilakukan dengan tahap-tahap berikut:

1. Pekerjaan peserta yang sudah dikoreksi dibagi ke dalam tiga kelompok atas dasar besarnya skor mentah. Ketiga kelompok itu ialah kelompok nilai tinggi, kelompok nilai sedang, dan kelompok nilai rendah. Penentuan jumlah kelompok atas dan bawah bisa berkisar antara 25%-33% tergantung pada besarnya jumlah sampel.
2. Melakukan penghitungan jumlah jawaban benar terhadap masing-masing butir soal pada setiap kelompok, menggunakan rumus berikut:

$$D = (T-R):N$$

dengan  $D$  = nilai diskriminasi,  $T$  = jawaban benar pada kelompok tinggi,  $R$  = jawaban benar pada kelompok rendah, dan  $N$  merupakan jumlah peserta tes.

Dengan asumsi bahwa  $T$  selalu lebih pandai daripada  $R$ , maka  $D$  negatif diinterpretasikan sebagai indeks yang menyimpang, yang mungkin disebabkan oleh susunan atau rumusan butir tes yang tidak wajar. Penetapan indeks diskriminasi dilakukan menurut kategorisasi berikut ini: 0.50 atau lebih = baik

$$0.20 - 0.50 = \text{sedang}$$

$$> 0.20 = \text{rendah}$$

$$0 = \text{tidak ada diskriminasi}$$

negatif = menyimpang

## **METODE**

### **Rancangan Penelitian**

Penelitian ini menggunakan rancangan kuantitatif. Hal ini sesuai dengan tujuan penelitian yang menggambarkan temuan dalam bentuk persentase, frekuensi, dan rerata. Dalam bagian tertentu, penelitian ini juga menggunakan uji statistik product moment dan KR-21. Namun penggunaan uji statistik tersebut ialah untuk memberikan penggambaran yang lebih jelas mengenai data yang sudah dijelaskan menggunakan frekuensi dan persentase. Instrumen penelitian berupa butir-butir soal UAS Bahasa Inggris untuk SMP tahun 2014/2015 yang disusun oleh Depdiknas. Sedangkan subjek penelitian ialah 58 siswa kelas III yang belajar di tahun terakhir di SMPN 1, SMPN 2, dan SMP Kesatuan Samarinda.

### **Subjek Penelitian**

Subjek penelitian berupa 58 lembar jawaban soal bahasa Inggris UAS SMP 2014/2015. Lembar jawaban diseleksi dari 120 lembar jawaban siswa yang tergabung dalam Rayon 1 UAS di Samarinda yaitu SMPN 1, SMPN 2, dan SMP Kesatuan. Jumlah 58 diperoleh sebagai berikut. Ketiga SMP di atas, dipilih 40 siswa sebagai sampel sehingga terkumpul 120 siswa. Ke-120 siswa kelas III yang akan mengikuti UAS tahun ajaran 2014/2015, diberi tes UAS tahun 2014/2015. Setelah lembar jawaban diseleksi dan jawaban yang rusak ditinggalkan, tercatat ada 58 lembar jawaban. Lembar ini dianggap sebagai sampel untuk analisis butir soal.

Ketiga sekolah tersebut dipilih sebagai sampel karena SMPN 1 dan SMPN 2 dianggap sebagai sekolah barometer di Samarinda. Sebagai sekolah unggulan, kedua SMP tersebut sering digunakan sebagai sumber kajian dan informasi. Selain itu, penyusun soal UAS untuk SMP pada tingkat provinsi Kalimantan Timur ialah guru dari kedua sekolah tersebut. Dengan demikian, sampel ini bisa dianggap mewakili kemampuan siswa sekolah negeri di Samarinda. Sedangkan SMP Kesatuan dipilih sebagai subjek yang mewakili SMP Swasta di Samarinda karena SMP Kesatuan ialah satu-satunya sekolah swasta di Samarinda yang statusnya disamakan. Dengan demikian sekolah ini berhak menyelenggarakan UAS sendiri.

### **Instrumen Penelitian**

Instrumen penelitian berupa satu set soal UAS Bahasa Inggris untuk SMP di Samarinda yang diujikan tahun 2014/2015. Instrumen terdiri dari 50 butir pertanyaan mencakup 45 butir soal pilihan ganda dengan empat pilihan dan 5 butir soal mengisi dengan jawaban yang sudah ditentukan. Butir soal ini dicetak oleh Depdiknas Jakarta, tetapi proses penyusunannya melibatkan guru-guru SMP di Kalimantan Timur dan Diknas Kalimantan Timur. Butir pertanyaan disusun dengan komposisi (1) reading 5 teks, (2) gramatika, kosa kata, dan (4) isian.

### **Teknik Pengumpulan Data**

Data penelitian dikumpulkan dalam tiga tahap. Pertama, lembar jawaban tes yang sudah diseleksi dan dianggap memenuhi syarat untuk dikaji diidentifikasi setiap butirnya dengan tanda benar dan salah. Identifikasi benar-salah ini menghasilkan rekaman jumlah butir yang dijawab benar dan yang dijawab salah dan skor mentah setiap siswa. Tahap kedua membagi lembar jawaban ke dalam kelompok atas dan kelompok bawah masing-masing 29 lembar. Tahap ini menghasilkan data untuk mengidentifikasi p dan D setiap butir soal. Setelah kedua tahap ini dilakukan, interpretasi juga dilakukan untuk melihat efektivitas option, validitas dan reliabilitas.

### **Teknik Analisis Data**

Analisis data dilakukan dalam lima kategori. Pertama, untuk mengetahui p, digunakan rumus:  $JPT/JPT \times 100\%$  (jumlah jawaban benar dibagi dengan jumlah peserta tes). Kedua, untuk mengetahui D, digunakan rumus:  $(T-R)/N$  (jawaban benar pada kelompok atas dikurangi dengan jawaban benar pada kelompok bawah, hasilnya dibagi dengan jumlah peserta tes). Ketiga, efektivitas option diperoleh dari jumlah penjawab benar untuk pilihan tertentu dibagi dengan jumlah peserta tes. Keempat, validitas kriteria diperoleh menggunakan rumus Product Moment by Carl Pearson. Kelima, koefisien reliabilitas diperoleh dengan rumus KR-21.

### **HASIL DAN BAHASAN**

Hasil penelitian terhadap lima variabel data disajikan sebagai berikut:

1. Analisis p butir soal menunjukkan bahwa dari 45 butir soal pilihan ganda 19 butir, atau 42.2% harus dibuang dan 26 butir lainnya atau 57.8% berada dalam p sedang. Analisis D menunjukkan bahwa dari 45 butir soal, 17 butir atau 37.8% daya pembedanya rendah, dan 28 atau 62.2% berdaya beda sedang. Jika p dan D digabung, hasil analisis menunjukkan bahwa di antara 45 butir soal, hanya 5 atau 11.1% yang bisa digunakan, 21 atau 46.7% harus direvisi dan 19 atau 42.2% harus dibuang.
2. Analisis distraktor menunjukkan hasil bahwa 7 soal atau 15.5% distraktor tidak berfungsi; 13 soal atau 28.9% memiliki distraktor sedang; dan 25 atau 55.6% memiliki distraktor sangat kurang. ( $p = .20 - .80$ ;  $D = .50$  ke atas; distraktor = 20% - 70%).
3. Koefisien validitas menghasilkan  $r = .346$ . Hasil ini berada dalam posisi rendah, sehingga seluruh butir tes tidak valid. Dengan kata lain, skor siswa tidak menunjukkan kemampuan menjawab soal bahasa Inggris secara benar. Koefisien sebesar .346 hanya cocok untuk soal-soal buatan guru yang digunakan untuk ulangan formatif.
4. Koefisien reliabilitas tes menghasilkan nilai 0.650 menunjukkan koefisien pada taraf sedang. Dengan demikian, butir tes ini bisa dipercaya sampai taraf sedang dalam memberikan informasi mengenai keajegan sekor yang dipereoleh siswa. Butir tes cukup reliabel.
5. Analisis terhadap lima butir soal isian pendek menunjukkan (1) tingkat kesulitan 100% antara .70-1 yang berarti sangat mudah, (2) empat butir memiliki daya pembeda di bawah 0,20 yang berarti hampir tidak berfungsi, dan (3) satu butir memiliki daya pembeda

negatif (-0,26) yang berarti siswa dalam kelompok rendah lebih pandai dibanding dengan siswa kelompok tinggi.

## SIMPULAN

Analisis terhadap soal tes UAS bahasa Inggris untuk SMP di Samarinda tahun 2014/2015 menunjukkan hasil yang tidak baik. Tingkat kesulitan soal berada dalam derajat sangat jelek dan sedang; daya pembeda soal tidak ada yang mencapai kriteria ideal (0.50); dan efektivitas option soal pilihan ganda rendah. Dari 45 butir soal, hanya 5 butir yang layak digunakan, sedangkan lainnya berkategori harus direvisi dan harus dibuang.

Analisis validitas hanya memungkinkan dilakukan terhadap validitas kriteria. Hasil analisis ini termasuk rendah (.346). Sedangkan analisis reliabilitas menunjukkan penyimpangan skor secara sempurna. Dengan kata lain, skor yang diperoleh siswa tidak ajeg karena kualitas soal tidak baik. Tingkat kesulitan yang rendah dan tingkat daya pembeda yang juga rendah berarti soal bisa dikerjakan secara benar baik oleh siswa yang pandai maupun siswa yang tidak pandai.

## DAFTAR RUJUKAN

- Butler, Christopher. 1985. *Statistics in Linguistics*. New York: Basil Blackwell.
- Departemen Pendidikan dan Kebudayaan. 1985. *Informasi UAS*. Jakarta.
- Departemen Pendidikan dan Kebudayaan. 1986. *Petunjuk Pelaksanaan Evaluasi Belajar Tahap Akhir Nasional*. Jakarta.
- Departemen Pendidikan dan Kebudayaan. 1987. *Pedoman Penulisan Soal Tes Hasil Belajar Bahasa Inggris UAS*. Jakarta.
- Djiwandono. M. Soenardi. 1996. *Tes Bahasa dalam Pengajaran*. Bandung: Penerbit ITB Bandung.
- Douglas, Brown. 1980. *Principles of Language Learning and Teaching*. Malwah, New Jersey: Prentice Hall Inc.
- Gronlund, Norman E. 1985. *Measurement and Evaluation in Teaching*. New York: Macmillan Publishing Company.
- Harris, David P. 1969. *Testing English as a Second Language*. New Delhi: McGraw Hill Publishing Co.
- Harrison, Andrew. 1983. *Language Testing Handbook*. Hongkong: Macmillan Publisher, Ltd.
- Kantor Departemen Pendidikan dan Kebudayaan. 1996. *Laporan UAS 2014/2015 SMP Se-Kodya Samarinda*. Samarinda: KanDepdiknas.
- Laporan Hasil UAS 2014/2015 di Rayon 1 Samarinda*. 1996. Samarinda: SMPN 1.
- Moore, Gary W. 1983. *Developing and Evaluating Educational Research*. Toronto: Little Brown Company.
- Weir, Cyrel. 1993. *Understanding and Developing Language Tests*. Singapore: Prentice Hall Inc.