**el-HiKMAH**
Jurnal Kajian dan Penelitian Pendidikan Islam
Vol. 10, No. 2, Desember 2016, h. 157-178
P-ISSN: 2086-3594
E-ISSN: 2527-4651

# PERBANDINGAN RELIABILITAS BUTIR SOAL BAHASA ARAB ANTARA PILIHAN GANDA DENGAN MENJODOHKAN PADA TES BUATAN GURU

**Siti Hajaroh***

**Abstrak:** Artikel ini mengemukakan perbedaan reliabilitas butir tes pilihan ganda dengan menjodohkan pada tes buatan guru Bahasa Arab. Penelitian ini merupakan penelitian kuantitatif dengan desain Quasi eksperimen. Sampel penelitian 200 santri yang diambil dengan teknik acak sederhana dari populasi 223 santri kelas I di Pondok Pesantren Walisongo Ngabar Ponorogo Jawa Timur tahun akademik 2009/2010. Teknik analisis data meliputi: pengujian persyaratan normalitas dengan *Uji Liliefors,* homogenitas varian dengan *Uji fisher*, serta pengujian hipotesis menggunakan uji-t. Hasil pengujian hipotesisi diperoleh untuk α 0,05, dk 10 - 1 = 9 adalah 2,262 sedangkan nilai $t_{hitung}$ = 1,368. Kriteria pengujian dua pihak apabila $-t_{a/2}<t_{hitung}<t_{a/2}$ maka $H_0$ diterima dan $H_1$ ditolak, sehingga -2,262 $<t_{hitung}<$ 2,262, artinya $H_0$ diterima dan menegaskan bahwa tidak terdapat perbedaan koefisien reliabilitas butir antara tes bentuk pilihan ganda dengan tes menjodohkan (*matching test*) pada mata pelajaran Bahasa Arab santri kelas I di Pondok Pesantren Walisongo Ngabar Ponorogo.

**Kata kunci**: perbandingan, reliabilities, pilihan ganda, menjodohkan (*matching test*).

## Pendahuluan

Dalam upaya peningkatan mutu pendidikan, kegiatan evaluasi mempunyai peran yang sangat penting. Sebab melalui evaluasi, pendidik akan mengetahui gambaran kemampuan para siswa yang dievaluasi. Dalam evaluasi selalu

*Fakultas Ilmu Tarbiyah dan Keguruan (FITK) IAIN Mataram. Email: hajaroh.saif@gmail.com

Copyright ©2016 el-Hikmah
Tersedia online di http:ejurnal.iainmataram.ac.id/index.php/el_hikmah

mengandung proses. Proses evaluasi harus tepat terhadap tujuan yang biasanya dinyatakan dalam bahasa perilaku. Dikarenakan tidak semua perilaku dapat dinyatakan dengan alat evaluasi yang sama, maka evaluasi menjadi salah satu hal yang sangat sulit dan menantang, yang harus disadari oleh para guru. Menurut Undang-Undang Republik Indonesia Nomor 20 tahun 2003 Tentang Sistem Pendidikan Nasional Pasal 57 ayat (1): Evaluasi dilakukan dalam rangka pengendalian mutu pendidikan secara Nasional sebagai bentuk akuntabilitas penyelenggaraan pendidikan kepada pihak-pihak yang berkepentingan, diantaranya terhadap peserta didik, lembaga dan program pendidikan.

Kegiatan evaluasi mempunyai hubungan dengan kegiatan pengukuran. Pengukuran adalah salah satu tahapan sangat penting dalam suatu proses penelitian ilmiah. Melalui pengukuran akan dihasilkan data penelitian, kemudian berdasarkan data penelitian tersebut dibuat penafsiran, kesimpulan dan implikasinya. Karena pencapaian perkembangan siswa perlu diukur terlebih dahulu, baik posisi siswa sebagai individu maupun dalam kelompok. Hal yang demikian perlu disadari oleh guru karena pada umumnya masing-masing siswa mempunyai kemampuan yang bervariasi. Ada siswa yang memiliki kemampuan cepat menangkap materi dan ada yang lambat. Guru dapat mengevaluasi pertumbuhan kemampuan siswa tersebut dengan mengetahui apa yang mereka kerjakan dari awal sampai akhir. Pencapaian belajar ini dapat dievaluasi melalui kegiatan pengukuran (*measurement*).

Mengukur pencapaian hasil belajar dapat melibatkan pengukuran secara kuantitatif yang menghasilkan data kuantitatif, misalnya tes dan skor, dan juga dapat mengukur data kualitatif yang menghasilkan deskripsi tentang subyek atau obyek yang diukur, misalnya rendah, medium, dan tinggi. Jadi kegiatan mengukur atau sering disebut pengukuran tidak lain adalah bagian evaluasi yang memiliki tujuan untuk menghasilkan data baik secara kuantitatif, maupun kualitatif (Sukardi, 2009: 2-3).

Dalam bidang pendidikan, pengukuran memegang peranan yang sangat penting. Sebab data hasil pengukuran dalam bidang

pendidikan memiliki arti penting baik bagi lembaga pendidikan, guru, peserta didik, orang tua, maupun masyarakat. Bagi guru pengukuran bermanfaat untuk membandingkan tingkat kemampuan siswa dengan siswa yang lain dalam kelompok yang diajarnya. Di sekolah pengukuran dilakukan guru untuk menaksir kemampuan atau prestasi siswa. Adapun alat yang digunakan untuk mengukur prestasi tersebut adalah berupa tes.

Tes kebahasaan sangat beragam, bergantung pada perbedaan tujuan, kepentingan, cara pemeriksaan, dan ruang lingkupnya. Dari segi tujuannya, tes kebahasaan dapat diklasifikasikan menjadi tiga, yaitu: tes pemerolehan atau tes prestasi *(achievement test, al-ikhtibâr al-tahshîlî*), tes profisiensi (*proficiency test, ikhtibâr al-ijâdah aw al-kafâah*), dan tes kesiapan berbahasa *(language aptitude test, ikhtibâr al-isti'dâd al-lughawî*) atau tes prekdisi (*predictive test, al-ikhtibâr al-tanabbu'*)(Hidayat, 2003:1).

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'bîr syafawî*). Sedangkan tes tulis dapat digunakan untuk menguji cabang-cabang ke-bahasa-Arab-an yang kurang cocok diujikan secara lisan, seperti: materi *nahwu, tarjamah tahrîriyyah* (terjemah tulis)*, insyâ'*, dan sebagainya (Hidayat, 2003:1).

Sementara itu, dari segi skoringnya, tes dapat dibagi menjadi dua, yaitu: tes *essay* atau tes subyektif dan tes obyektif. *Pertama,* 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. *Kedua,* tes yang itemnya 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-shawâbwa al-khatha'*), mencari pasangan (*al-muzâwajah, matching*), dan melengkapi isian (*al-takmilah, completion*) dengan jawaban yang bersifat tertutup (Hidayat, 2003:1). Kompetensi yang diukur dalam Bahasa Arab mencakup *nahwu*, *sharf*, dan *mufradat*. *Nahwu* adalah ilmu yang mempelajari tentang susunan kata dalam hal kedudukan akhir kata dalam sebuah kalimat. Tujuan dari pembelajaran *nahwu* adalah kemampuan menganalisis unsur-unsur kalimat secara analisis bahasa di mana dapat dipahami bagian-bagian yang terdapat dalam kalimat sehingga dapat dijelaskan unsur-unsur pembentuk suatu kalimat (Sulaiman Fayyad, 2003:13). Sementara *sharf*, secara umum membahas dua bagian: a) intrapolasi kata Bahasa Arab ke dalam ragam bentuk untuk memperoleh bermacam arti kata yang diinginkan, dan b) mengubah bentuk kata dari bentuk asalnya tanpa mengubah arti kata yang dikandung (Al-Sayyid, 1972: 17-19) Sedangkan *mufradat*, merupakan bagian dari kajian *sharf* yang oleh sebagian ahli bahasa diungkapkan dengan istilah kata. Pembahasan kata dalam Bahasa Arab meliputi dua segi; a) penguasaan arti dari setiap kata dan pembedaan arti yang sesungguhnya *(al-makna al-hakiki)* dengan arti kiasan *(al-makna al-majazi)*, b) pengucapan, yakni ketepatan bunyi (harakat) setiap huruf yang ada dalam kata (al-Sayyid, 1972: 17-19)

Adapun fungsi evaluasi dalam dunia pendidikan tidak dapat dilepaskan dari tujuan evaluasi itu sendiri. Di dalam batasan tentang evaluasi pendidikan yang telah dijelaskan sebelumnya, tersirat bahwa tujuan evaluasi pendidikan adalah untuk mendapat data pembuktian yang akan menunjukkan sampai sejauh mana tingkat kemampuan dan keberhasilan siswa dalam pencapaian tujuan-tujuan kurikuler. Disamping itu juga dapat digunakan oleh guru-guru atau pengawas pendidikan untuk mengukur atau menilai

sampai sejauh mana keefektifan pengalaman-pengalaman mengajar, kegiatan-kegiatan belajar, dan metode-metode mengajar yang digunakan. Dengan demikian, dapat dikatakan betapa pentingnya peranan dan fungsi evaluasi dalam proses belajar-mengajar. Secara umum evaluasi sebagai tindakan atau proses setidaknya memiliki tiga macam fungsi pokok, yaitu: 1). Mengukur kemajuan, 2). Menunjang penyusunan rencana, 3). Memperbaiki dan melakukan perbaikan kembali (Sudijono, 2009:8).

Mengenai betapa pentingnya sebuah evaluasi dalam kegiatan pembelajaran, Mehrens dan Lehmann dalam Djaali mengutip suatu ungkapan yang berbunyi "*to teach without testing is unthinkable*" (mengajar tanpa melakukan tes tidak masuk akal) (Djaali dan Mulyono, 2008:2). Maka, dapat disimpulkan bahwa evaluasi adalah sebuah proses tertentu yang digunakan untuk menafsirkan perkembangan peserta didik baik positif ataupun negatif mengenai kurikulum tertentu sebagaimana yang telah ditetapkan, dengan tujuan untuk membuat sebuah keputusan. Jadi, setiap kegiatan evaluasi merupakan suatu proses yang direncanakan untuk memperoleh informasi atau data, yang kemudian data tersebut akan digunakan untuk membuat sebuah keputusan. Sudah barang tentu bahwa informasi atau data yang dikumpulkan harus sesuai dan mendukung tujuan dari evaluasi yang telah direncanakan tersebut.

Secara teoritis, kemampuan siswa dalam satu kelas merupakan kelompok yang sifatnya heterogen. Dengan demikian, maka apabila dikenai sebuah tes akan tercermin hasilnya dalam suatu kurva normal. Sebagian besar berada pada kelompok sedang, sebagian kecil berada pada kelompok tinggi dan sebagian kecil lainnya berada pada kelompok rendah. Apabila keadaan dimana nilai-nilai *testee* membentuk kurva a-simetrik, maka tentu ada sesuatu yang kurang beres sehingga perlu diadakannya antisipasi.

Salah satu cara mengantisipasi keadaan yang tidak normal itu adalah dengan jalan penganalisisan terhadap tes hasil belajar yang telah dijadikan alat pengukur dalam rangka mengukur keberhasilan belajar dari para *testee* tersebut. Disini *tester* perlu mengadakan penelusuran dengan cermat terhadap butir-butir soal atau item yang

merupakan bagian yang tak terpisahkan dari tes hasil belajar sebagai suatu totalitas. Tujuan dari penelusuran itu adalah untuk mengetahui apakah butir-butir atau item soal yang membangun tes hasil belajar sudah dapat menjalankan fungsinya sebagai alat pengukur hasil belajar yang memadai ataukah belum. Adanya identifikasi itu diharapkan akan menghasilkan informasi berharga, yang pada dasarnya akan menjadi umpan balik (*feed back*) guna melakukan perbaikan, pembenahan, penyempurnaan kembali terhadap butir-butir item yang telah dikeluarkan dalam tes hasil belajar, sehingga pada masa yang akan datang tes hasil belajar yang dirancang oleh *testee* benar-benar dapat menjalankan fungsinya sebagai alat ukur hasil belajar yang memiliki kualitas yang tinggi. Kegiatan identifikasi atau penelusuran itu dikenal dengan analisis butir soal atau analisis item (*item analysis*).

Menurut Sudjana analisis butir soal adalah pengkajian pertanyaan tes agar diperolah perangkat pertanyaan yang memiliki kualitas yang memadai (Sudjana, 2009:135). Adapun analisis soal ini dilakukan untuk mengetahui berfungsi tidaknya sebuah soal. Analisis pada umumnya dilakukan melalui dua cara, yaitu; analisis kualitatif (q*ualitatif control*) dan analisis kuantitatif (*quantitatif control*). Analisis kualitatif sering pula dinamakan sebagai validitas logis (*logical validity*) yang dilakukan sebelum soal digunakan untuk melihat berfungsi tidaknya sebuah soal. Analisis soal secara kuantitatif sering pula dinamakan sebagai validitas empiris (*empiric validity*) yang dilakukan untuk melihat lebih berfungsi tidanya sebuah soal setelah soal itu diujicobakan kepada sampel yang representatif.

Analisis soal kuantitatif menekankan pada analisis karakteristik internal tes melalui data yang diperoleh secara empiris. Karakteristik internal secara kuantitatif dimaksudkan meliputi parameter soal tingkat kesukaran, daya beda, dan reliabilitas. Khusus soal-soal pilihan ganda, dua tambahan parameter yaitu dilihat dari peluang untuk menebak atau menjawab soal benar dan berfungsi tidaknya pilihan jawaban, yaitu penyebaran semua

alternatif jawaban dari subyek-subyek yang dites (Sumarna Surapratama, 2004: 10).

Adapun salah satu tujuan diadakanya analisis menurut Surapranata adalah untuk meningkatkan kualitas soal, yaitu apakah suatu soal, 1). *Dapat diterima*, karena telah didukung oleh data statistik yang memadai, 2). *Diperbaiki*, karena terbukti terdapat beberapa kelemahan atau bahkan, 3). *Tidak digunakan sama sekali*, karena terbukti secara empiris tidak berfungsi sama sekali (Sumarna Surapratama, 2004: 10).

Sehingga dari beberapa pernyataan di atas dapat disimpulkan bahwa analisis butir adalah kegiatan menganalisis item-item tes dengan tujuan untuk mengatahui kualitas butir-butir soal yang telah dibuat. Adapun penganalisisan dari pada butir-butir soal sebagaimana dijelaskan di atas salah satunya adalah reliabilitas tes.

Reliabilitas adalah karakter lain dari evaluasi, Reliabilitas juga dapat diartikan sama dengan konsistensi atau keajegan. Suatu instrumen evalusi, dikatakan mempunyai nilai reliabilitas tinggi, apabila tes yang dibuat mempunyai hasil yang konsisten dalam mengukur apa yang hendak diukur. Reliabilitas memberikan konsistensi yang membut terpenuhinya syarat utama yaitu validnya suatu hasil Skor instrumen. Disamping itu, reliabilitas juga menunjukkan gambaran praktis yang dapat diklasifikasian berkaitan erat dengan syarat ketiga, yaitu kebermanfaatan (*usability*). Ini berarti semakin tinggi reliabel suatu tes, semakin yakin kita dapat menyatakan bahwa dalam hasil suatu tes mempunyai hasil yang sama dan bisa dipakai di suatu tempat sekolah, ketika dilakukan tes kembali.

Reliabilitas suatu tes adalah seberapa besar derajat tes mengukur secara konsisten sasaran yang diukur. Reliabilitas dinyatakan dalam bentuk angka, biasanya sebagai koefisien. Koefisien tinggi berarti reliabilitas tinggi (Sukadji, 2010:3). Menurut McMillan reliabilitas adalah derajat ketepatan terh terhadap skor yang diperoleh dari pengukuran (McMillan, 2008:35). Sependapat dengan McMillan, Naga mendefinisikan reliabilitas pada hakikatnya adalah tingkat kepercayaan terhadap skor atau tingkat kecocokan skor dengan

skor sesungguhnya. Reliabilitas dicapai melalui tingkat kecocokan di antara skor pada lebih dari satu pengukuran (Naga, 2008:160). Sedangkan menurut Arikunto, reliabilitas berhubungan dengan masalah kepercayaan. Suatu tes dapat dikatakan mempunyai taraf kepercayaan jika tes tersebut dapat memberikan hasil yang tetap. Maka pengertian reliabilitas tes, berhubungan dengan masalah ketetapan hasil tes (Arikunto, 1999:86).

Gronlund mengemukakan bahwa keterandalan menunjuk kepada konsistensi (*keajegan*) pengukuran yakni bagaimanakah keajegan skor tes atau hasil evaluasi yang berasal dari pengukuran yang satu dengan yang lain (Gronlund, 1985:81). Menurut Anastasi dan Urbina, reliabilitas adalah kestabilan skor yang diperoleh dari orang yang sama pada situasi yang berbeda (Anastasi, 1997:63). Sehingga dengan kata lain, keterandalan dapat kita artikan sebagai tingkat kepercayaan keajegan hasil evaluasi yang diperoleh dari suatu instrumen evaluasi. Hal ini sesuai dengan pendapat Wiersma dan Jurs yang menyatakan bahwa reliabilitas pengukuran adalah hasil tetap (*consistency*), ketetapan dalam mengukur instrumen (*test*) yang digunakan untuk mengukur (Wiersma, William, dan Jurs, 2001:155). Reliabilitas alat ukur tidak dapat diketahui dengan pasti tetapi dapat diperkirakan.

Sehingga dapat disimpulkan bahwa reliabilitas merupakan derajat konsistensi skor yang dicapai oleh seseorag dalam hal ini adalah orang yang sama, ketika mereka diuji-ulang dengan tes yang sama pada kesempatan yang berbeda, atau dengan seperangkat butir-butir ekuivalen yang berbeda, atau di bawah kondisi pengujian yang berbeda.

Menurut Gronlund 4 faktor yang mempengaruhi keterandalan yaitu:

1. Panjag tes (*Length of test*). Pada umumnya lebih banyak butir tes lebih tinggi reliabilitas evaluasi. Hal ini karena makin banyak soal tes makin banyak sampel yang diukur, proporsi jawaban benar makin banyak, dengan demikian faktor tebakan (*guessing*) makin rendah karena pengertian tes dilakukan dengan tidak banyak menebak, maka keterandalan evaluasi semakin tinggi.

2. Sebaran skor (*spread of score*). Koefisien reliabilitas secara langsung dipengaruhi oleh sebaran skor dalam kelompok tercoba. Dengan kata lain besarnya sebaran skoran membuat perkiraan reliabilitas yang lebih tinggi akan menjadi kenyataan. Karena koefisien yang lebih besar dihasilkan pada satu orang perorang tetap pada posisi yang relatif sama dalam satu kelompok dari satu pengujian ke pengujian yang lainnya. Itu berarti selisih yang dimungkinkan dari perubahan posisi dalam kelompok juga menyumbang memperbesar koefisien reliabilitas.
3. Tingkat kesulitan tes (*difficulty of test*). Tes acuan Norma yang paling mudah atau yang paling sukar untuk anggota-anggota kelompok yang mengerjakan, cenderung menghasilkan skor tes keterandalan yang rendah. Ini disebabkan antara hasil tes yang mudah dan yang sulit keduanya dalam satu sebaran skor yang terbatas. Untuk tes yang mudah skorakan berada bersama-sama pada bagian atas dan akhir skala penilaian dan seliknya. Tingkat kesulitan tes yang ideal untuk meningkatkan reliabilitas adalah tes yang menghasilkan sebaran skor berbentuk kurva normal.

Objektifitas (*objektifity*). Objektifitas tes menunjuk pada tingkat skor kemampuan yang sama (*yang dimiliki siswa yang satu dengan siswa yang lain*) memperoleh hasil yang sama dalam mengerjakan tes. Dengan kata lain apabila siswa yang memiliki tingkat kemampuan yang sama dengan tingkat kemampuan siswa yang lain maka dipastikan akan memperoleh hasil yang sama saat mengerjakan tes yang sama. Objetifitas prosedur tes yang tinggi akan menghasilkan keterandalan hasil tes yang tidak dipengaruhi oleh prosedur penskoran (Gronlund, 1985:100).

Sependapat dengan pendapat di atas McMillan juga memberikan pendapatnya bahwa reliabilitas dipengaruhi oleh sebaran skor, jumlah item tes, tingkat kesukaran pada butir-buitr item dalam tes, Kualitas item tes, objektifitas, perbedaan niliai antar individu. Tidak reliabelnya suatu tes pada prinsipnya dapat dikatakan sia-sialah tes tersebut, karena jika dilakukan pengetesan kembali hasilnya akan berbeda. Reliabilitas suatu tes pada umumnya diekspresikan secara numerik dalam bentuk koefisien yang besarnya

*-1<0<1*. Koefisisen tinggi menunjukkan reliabilitas tinggi, sebaliknya jika koefisien rendah maka reliabilitas tes juga rendah (McMillan, 2008:47-50).

Dalam suatu kenyataan tes yang mempunyai koefisien reliabilitas sempurna adalah tidak ada. Karena skor itu kemungkinan besar bervariasi, yang disebabkan oleh terjadinya kesalahan pengukuran yang berasal dari bermacam-macam sumber.Kesalahan pengukuran dapat disebabkan oleh beberapa faktor diantaranya karakteristik tes itu sendiri, kondisi pelaksanaan tes yang tidak mengikuti aturan baku, tes item yang meragukan dan peserta tes langsung mengikuti, status peserta yang mengikuti tes, misalnya seorang yang sedang lelah, atau mempunyai problem pribadi, peserta tes yang mempunyai motivasi rendah, atau kombinasi dari semua gejala tersebut (McMillan, 2008:44).

Reliabilitas yang tinggi menunjukkan bahwa sumber-sumber kesalahan telah dihilangkan sebanyak mungkin. Perhitungan reliabilitas pada umumnya lebih mudah dibanding dengan validitas. Hal ini terjadi karena dalam menentukan koefisien korelasi, peneliti tidak lagi memikirkan substansi dalam tes.

Ada beberapa type reliabilitas dalam tes yang sering digunakan dalam kegiatan evaluasi dan masing-masing reliabilitas mempunyai konsistensi yang berbeda-beda. Beberapa type reliabilitas itu antara lain; tes-retes, ekivalen, belah dua dan Kuder Richardson (*K-R*) (McMillan, 2008:44).

**Tabel 1. Macam-Macam Reliabilitas dan Prosedur Pelaksanaan Pengukuran Reliabilitas**

| Metode | Type Reliabilitas Pengukuran | Prosedur |
|---|---|---|
| Tes-retes | Mengukur stabilitas | Memberikan tes yang sama, dua kali pada group yang sama dengan jeda waktu diantara dua tes, misalnya 7 hari – 1 bulan |
| Ekivalen | Mengukur ekivalen Mengukur stabilitas dan ekivalen | Memberi dua bentuk tes pada gruop yang sama dalam waktu yang berdekatan. Dua bentuk tes diberikan pada dua gruop siswa dengan menambah interval antara bentuk |

| | | |
|---|---|---|
| Belah dua | Mengukur kesulitan internal | Diberikan tes sekali separuh skor tes; diperoleh korelasi antara separuh untuk menempatkan semua tes dengan Sperman Brown |
| Kuder Richardson | Mengukur konsistensi internal | Diberikan ekor tes sekali, tes skor merupakan skor total dengan menggunakan formula Kuder Richardson |

Konsistensi internal tergantung pada interkorelasi butir tes, yang disebut homogenitas. Rumus statistik terbaik yang digunakan untuk menentukan koefisien reliabilitas konsistensi internal adalah Alpha Cronbach dan Kuder-Richardson (KR-20 dan KR-21).

Reliabilitas berhubungan dengan konsistensi hasil pengukuran. Reliabiltas dipengaruhi oleh cakupan instrumen penilaian. Misalnya, suatu instrumen tes tertentu yang mencakup sasaran belajar dan butir yang terbatas memiliki reliabilitas yang lebih rendah dibanding dengan tes yang mencakup sasaran belajar yang lebih luas dengan jumlah butir yang lebih banyak.

Pada mata pelajaran Bahasa Arab khususnya, kedua bentuk tes obyektif yaitu bentuk pilihan ganda dan mencari pasangan pada dasarnya hampir sama, karena pada soal pilihan ganda memiliki alternatif jawabannya lebih sedikit akan memiliki pengecoh atau *distractor,* artinya dalam soal pilihan ganda hanya ada satu alternatif jawaban yang paling benar dan yang lain adalah pengecoh yang berfungsi untuk mengalihkan perhatian peserta tes. Sedangkan pada bentuk tes menjodohkan memiliki alternatif jawaban yang lebih banyak sehingga kesalahan menjodohkan dalam satu pasangan akan mengakibatkan kesalahan dalam menjodohkan pasangan yang lain. Sehingga untuk mengetahui tingkat kesukaran, daya beda dan reliabilitas kedua bentuk tes ini sangatlah sulit khususnya pada mata pelajaran Bahasa Arab. Selama ini item tes pilihan ganda merupakan jenis tes obyektif yang paling banyak digunakan oleh para guru. Karena beberapa penelitan membuktikan bahwa tes ini mampu mengukur pengetahuan yang luas dengan tingkat domain yang bervariasi dan item pilihan ganda memiliki semua persyaratan sebagai tes yang baik, yakni dilihat dari validitas, reliabilitas, dan

daya pembeda antara siswa yang berhasil dengan siswa yang tidak berhasil.

Untuk itu perlu adanya analisis untuk membuktikan masalah tersebut, sebab tes atau instrumen – instrumen Bahasa Arab yang dipakai untuk mengukur kemampuan siswa khususnya pada ujian pondok belum pernah dianalisis sehingga sampai saat ini kualitas butir soal tersebut belum diketahui. Untuk itu perlu adanya penganalisaan butir-butir soal tersebut.

## Metode Penelitian

Desain yang digunakan dalam penelitian ini adalah dengan menggunakan metode *Quasi Eksperimen*. Desain ini mempunyai kelompok kontrol, akan tetapi tidak berfungsi sepenuhnya untuk mengontrol variabel-variabel luar yang mempengaruhi pelaksanaan eksperimen. Quasi experimen digunakan karena pada kenyataanya sulit mendapatkan kelompok kontrol yang digunakan untuk penelitian (Sugiyono, 2008:77). Adapun rancangan penelitian yang digunakan adalah dengan menggunakan rancangan satu kelompok dimana subyek diacak dengan cara memberikan tes akhir saja atau dikenal dengan (*The Posttest-Only Design with nonequivalent Group*)"( Thomas D. Cook dan Donald T. Campbell, 1979:13).

| | |
|---|---|
| X | o |
| | o |

**Tabel 2. Matrik Desain Quasi Esperiment**

| Soal Analisis / Bentuk Tes | Reliabilitas |
|---|---|
| Pilihan Ganda | R-PG |
| Matching Tes | R-M |

Penelitian ini merupakan penelitian kuantitatif. Instrumen yang digunakan adalah instrumen tes hasil belajar Bahasa Arab. Bentuk intrumennya berupa pilihan ganda dan menjodohkan (*matching test*),

kemudian skor dari tiap-tiap butir tersebut dianalisis dengan menggunakan analisis butir dengan tujuan untuk menentukan koefisien reliabilitas yang selanjutnya akan dilakukan analisis pengujian hipotesis penelitian.

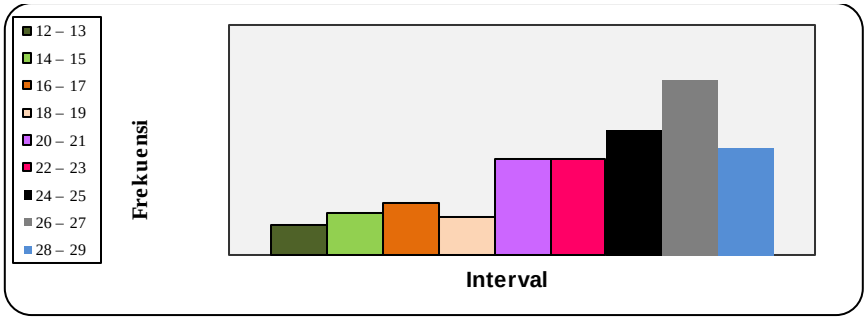
Populasi penelitian adalah seluruh santri kelas I dan I intensif TMI (*Tarbiyatul Muallimin Al-Islamiyah)* dan TMT (*Tarbiyatul mu'alimat Al-Islamiyah*) Pondok Pesantren Walisongo Ngabar Ponorogo Jawa Timur tahun akademik 2009/2010 yang terdiri dari 223 santri. Dengan teknik acak sederhana (*simple random sampling*) sampel penelitian mengambil 200 santri Mu'alimin dan Mu'alimat.

Teknik untuk menganalisis data meliputi beberapa tahap, yaitu: melakukan uji persyaratan analisis yang meliputi uji normalitas (*Uji Liliefors*) dan uji homogenitas varian (*Uji fisher*) karena hanya membandingkan dua kelompok data, menguji hipotesis penelitian yakni membandingkan rata-rata menggunakan uji-t.

## Hasil dan Pembahasan

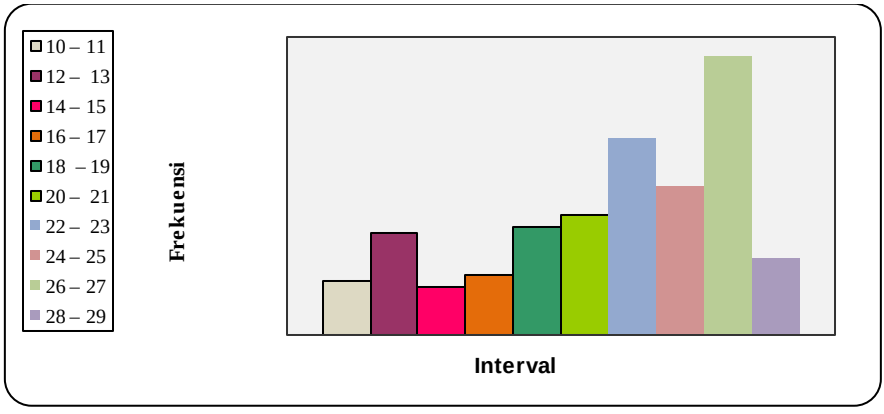
Hasil penelitian yang dilakukan pada mata pelajaran Bahasa Arab, kemudian diperoleh skor tes baik pilihan ganda dan *matching test* dengan tujuan untuk membandingkan rerata reliabilitas skor. Jumlah siswa yang dijadikan sampel pada kelas dalam rangka untuk pengambilan skor tes pilihan ganda adalah berjumlah 200 santri. Hasil perhitungan data penelitian mengenai skor pilihan ganda pada mata pelajaran Bahasa Arab, memiliki rentang nilai 12 sampai 29, dengan nilai rata-rata (*mean*) sebesar 22,93; *modus* 26,338 dan *median* 23,924.

Deskripsi data skor tes pilihan santri pada mata pelajaran Bahasa Arab ditunjukkan pada grafik dibawah ini:




**Gambar 1. Histogram Data Tes Pilihan Ganda**

Jumlah siswa yang dijadikan sampel pada kelas dalam rangka pengambilan skor pilihan ganda adalah berjumlah 200 santri. Hasil perhitungan data penelitian mengenai skor tes menjodohkan (*matchingtest*) pada mata pelajaran Bahasa Arab rentang nilai 10 sampai 28, dengan nilai rata-rata (*mean*) sebesar 21,47; *modus* 26,286 dan *median* 22,59.



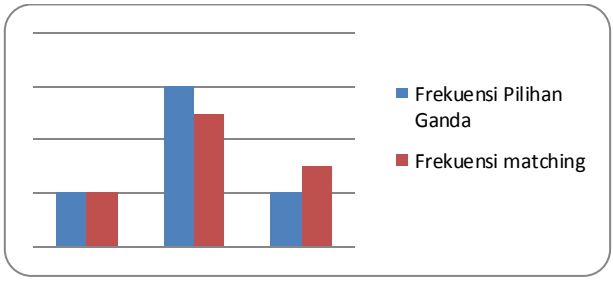

**Gambar 2. Histogram Data Bentuk *Matching Tes* (Menjodohkan)**

Data reliabilitas tes diperoleh dari perhitungan analisis terhadap skor santri yang berasal dari dua bentuk tes yaitu pilihan ganda dengan menjodohkan (*matching test)*. Adapun rincian data reliabilitas tes dengan melakukan acak (*random*) terhadap 200 santri, serta melakukan pengulangan uji coba sebanyak 10 kali (n=10) sebagai berikut:

**Tabel 3. Data Reliabilitas pada Soal Pilihan Ganda dengan *Matching***

| Uji Reliabilitas ke- | Bentuk Tes | |
|---|---|---|
| | Pilihan Ganda | Matching |
| 1 | 0.8311 | 0.8505 |
| 2 | 0.7741 | 0.8747 |
| 3 | 0.8009 | 0.878 |
| 4 | 0.5768 | 0.8476 |
| 5 | 0.8549 | 0.8485 |
| 6 | 0.8466 | 0.8692 |
| 7 | 0.8652 | 0.8392 |
| 8 | 0.8268 | 0.7965 |
| 9 | 0.8018 | 0.7852 |
| 10 | 0.8559 | 0.8311 |

Berdasarkan pada hasil analisis data di atas, maka frekuensi reliabilitas tes pada dua bentuk tes yaitu pilihan ganda dengan menjodohkan (*matching test)* dapat digambarkan pada histogram berikut:




**Gambar 3. Histogram Tingkat Reliabilitas Soal pada Tes Pilihan Ganda dengan Tes Menjodohkan (*Matching test).***

Gambar histogram di atas menunjukkan bahwa pada interval 0.579-0,7 tes soal 97 pilihan ganda memiliki frekuensi yang sama dengan test menjodohkan (*matching test)*, sedangkan pada interval antara 0.801 – 0.855 frekuensi reliabilitas soal pilihan gandalebih tinggi di banding dengan tes soal menjodohkan (*matching test)* dan

pada 0,856– 0,90 frekuensi reliabilitas soal menjodohkan (*matching test)* lebih tinggi di banding dengan tes soal pilihan ganda.

Uji normalitas dilakukan terhadap skor atau nilai tes pilihan ganda dan tes matching menggunakan uji *Lilliefors*. Dalam penelitian ini terdapat dua variabel data yaitu: (1) tes pilihan ganda dan (2) tes menjodohkan (*matching test)*, (2) Hasil perhitungan dan uji signifikansi indeks normalitas (harga L) secara keseluruhan dapat disajikan pada tabel berikut:

**Tabel 4. Rangkuman Hasil Uji Normalitas Data Skor Butir Tes Bentuk Pilihan Ganda dengan Tes Menjodohkan (*Matching test)***

| Bentuk Tes | n | Bentuk Tes | $L_{hitung}$ | $L_{tabel}$ | Kriteria Pengujian |
|---|---|---|---|---|---|
| Pilihan Ganda | 30 | Pilihan Ganda | 0,0930 | 0,886 | $L_{hitung} < L_{tabel}$ |
| Matching | 30 | Matching | 0,1058 | 0,886 | $L_{hitung} < L_{tabel}$ |

Tabel di atas menunjukkan bahwa harga *Lilliefors* hitung ($L_h$) pada masing-masing variabel lebih kecil dari harga *Lilliefors* tabel ($L_t$). Dengan demikian dapat disimpulkan bahwa sampel penelitian berasal dari populasi yang berdistribusi normal, sehingga analisis statistik parametrik dapat digunakan dalam analisis penelitian.

Uji Homogenitas Varians dilakukan dengan menggunakan *Uji Fisher*. Perhitungan pengujian kedua kelompok pada taraf signifikansi $\alpha = 0{,}05$ disajikan pada tabel berikut:

**Tabel 5. Rangkuman Hasil Uji *Fisher* Data Skor Butir Tes Bentuk Pilihan Ganda dengan Tes Menjodohkan (*Matching test)***

| Kelompok | N | Db | $s^2$ |
|---|---|---|---|
| Pilihan Ganda | 30 | 29 | 27,286 |
| Matching | 30 | 29 | 21,070 |

Dari hasil perhitungan yang diperoleh, kemudian dibandingkan dengan Ftabel pada db pembilang = 29 dan db penyebut = 29 Ftabel=F(0,05)(199;199)=1,28 dan Ftabel=F(0,01)(199;199)=1,39. Karena Fhitung lebih kecil dari Ftabel maka H0 diterima. Jadi

kedua distribusi populasi adalah mempunyai varians yang sama atau homogen.

Untuk mengetahui hasil penelitian yang membuktikan hipotesis ini maka terlebih dahulu dilakukan analisis dan penghitungan reliabilitas tes dengan menggunakan KR-20 dengan mengacak 75 siswa dari 200 santri yang dijadikan sampel untuk dua bentuk tes yaitu pilihan ganda dengan matching.

Untuk menentukan uji-t yang akan dipilih dalam menentukan pengujian hipotesis, maka perlu diuji dulu varians kedua sampel homogen atau tidak homogen. Dan berdasarkan hasil analisis dengan menggunakan uji F, maka menunjukkan bahwa Fhit= 74,481 dan Ftab = 3,18 pada taraf signifikan α = 0,05, maka Fhit>Ftab, artinya varians reliabilitas kedua bentuk tes yaitu pilihan ganda dengan matching adalah tidak homogen.

Selanjutnya hasil pengujian hipotesisi dengan menggunakan uji-t menunjukkan α=0,05 dengan dk=10-1=9 adalah 2,262 sedangkan nilai thitung = 1,368 dengan kriteria pengujian dua pihak dimana – ta/2<thitung<ta/2 maka H0 diterima dan H1 ditolak, sehingga -2,262 < thitung < 2,262, artinya H0 diterima pada taraf signifikan α=0,05. Dengan demikian dapat disimpulkan bahwa tidak terdapat perbedaan koefisien reliabilitas butir antara tes bentuk pilihan ganda dengan tes menjodohkan (matching test). Pada mata pelajaran Bahasa Arab Santri Pada kelas I di Pondok Pesantren Walisongo Ngabar Ponorogo.

Berdasarkna hasil analisis yang diperoleh di atas, pembuktian bahwa secara teoritik yang menyatakan bahwa bentuk pilihan ganda dan mencari pasangan pada dasarnya sama, karena pada soal pilihan ganda memiliki alternatif jawabannya lebih sedikit akan memiliki pengecoh atau distractor, artinya dalam soal pilihan ganda hanya ada satu alternatif jawaban yang paling benar dan yang lain adalah pengecoh yang berfungsi untuk mengalihkan perhatian peserta tes. Sedangkan pada bentuk tes menjodohkan memiliki alternatif jawaban yang lebih banyak sehingga kesalahan menjodohkan dalam satu pasangan akan mengakibatkan kesalahan dalam menjodohkan pasangan yang lain artinya dalam menjawab soal pada kedua bentuk

tes ini sama-sama memerlukan pemikiran yang kompleks. Dalam hal ini siswa tidak boleh menebak karena pada kedua bentuk tes ini siswa pempunyai peluang yang sama dalam menebak alternatif jawaban yang dipilihnya, karena ketelitian menjawab sangat diperlukan.

## Catatan Akhir

Sesuai tujuan dan permasalahan yang telah dirumuskan bahwa tujuan dari penelitian ini adalah menganalisis butir soal Bahasa Arab khususnya kelas I pada ujian pondok yang ditinjau dari reliabilitas soal pada tes pilihan ganda dan tes menjodohkan (*matching test)*. Berdasarkan hasil pengujian hipotesis yaitu membandingkan rerata reliabilitas antara tes bentuk pilihan ganda dengan matching, maka kesimpulannya tidak terdapat perbedaan koefisien reliabilitas pada bentuk tes pilihan ganda koefisien reliabilitas tes pada soal bentuk *matching*. Ini membuktikan sulitnya untuk membedakan reliabilitas antara tes pilihan ganda dengan menjodohkan (*matching test*) dimana kedua bentuk tes ini sama-sama memberikan peluang responden dalam menebak jawaban.

Pertama, Keberhasilan proses belajar mengajar tidak dapat dipantau tanpa adanya evaluasi hasil belajar. Akan tetapi alat evaluasi yang baik akan mungkin jika alat evaluasinya juga baik, maka guru dituntut untuk menguasai cara dan kaidah dalam menyusun tes yang baik. Untuk mengetahui apakah alat evaluasi tersebut baik atau tidak, maka perlu adanya analisis butir. Melalui analisis butir soal guru akan mendapatkan informasi untuk memberikan umpan balik baik kepada siswa maupun pendidik itu sendiri hasil dan dapat dipakai untuk mengupayakan perbaikan butir soal tersebut.

Kedua, dalam memilih bentuk soal, pendidik hendaknya menyesuaikan dengan karakteristik bidang studi dan mengetahui kualitas butir soal yang akan diujikan. Sebab dengan beragamnya bentuk soal obyektif tentunya masing-masing mempuyai kelebihan dan kekurangan masing-masing. Sebagai pendidik hendaknya tidak memberikan *jagement* bahwa bentuk soal pilihan ganda adalah lebih

baik jika dibanding bentuk soal obyektif yang lain. Sebelum soal diujikan kepada responden hendaknya butir-butir soal tersebut dianalisis terlebih dahulu dalam rangka untuk mengetahui mengetahui kualitas butir soal yang akan diujikan.

Ketiga, banyaknya hasil penelitian yang menyebutkan bahwa item pilihan ganda memiliki semua persyaratan sebagai tes yang baik, yakni dilihat dari validitas, reliabilitas, dan daya pembeda antara siswa yang berhasil dengan siswa yang tidak berhasil, tidak semuanya benar jadi sebelum memberikan keputusan dalam membuat soal hendaknya melakukan penyelidikan terlebih dahalu terhadap butir-butir soal yang akan diujikan. Dengan hasil penelitian ini diharapkan dapat memberikan masukan yang baik bagi dunia pendidikan terutama pendidik dalam rangka peningkatan mutu di sekolah-sekolah seluruh tanah air.

## Referensi


Abdulhakimmuh.wordpress.com/*artikel tentangSasaran Tes Unsur Bahasa Arab, Bentuk dan Susunannya, diakses* tanggal 29 Juni 2010.

Al-Sayyid, Amin Ali, *Fi Ilmi Al-Sharf* , Kairo: Dar-Al-ma'arif, 1972.

Aiken, dalam *Surapranata, Analisis validitas, Reliabilitas, dan Interpretasi Hasil Tes,* Bandung: Remaja Rosdakarya, 2004.

Anastasi, Anne, *Psychological Testing,* New York: Mac Millan Publishing Company, 1988.

Arikuto, Suharsimi, *Dasar-dasar Evaluasi Pendidikan,* Jakarta :Bumi Aksara, 1999.

Bloom, Benjamin S., *Taxonomy of EducatinObjectivees, Hanbook I: Cognitive Domaian*, New York; David Mc. Kay Company, Inc, 1971.

Blood Don F., and Budd, William C., *Educational Measurement and Evaluation* New York: Harperand Row, 1972.

Cook, Thomas D & Campbell, Donald T, *Quasi-Experimentation: Design & Analysis Issues for Field Settings., Alih bahasa*

(*Ringkasan buku*)*oleh DickyHastjarjo, th. 2008,* Boston: Houghton Mifflin Company, 1979.

Cronbach, Lle, *Essential of psychological testing,* New York: Harper & Row, 1984.

Crocker, L, *Item analysis*. Dalam Alkin M.C. (Eds.), Encyclopedia of Educational research, New York: Macmillan Library reference USA, 1992.

Dahlan, Djuwariyah, *Metode Belajar Mengajar Bahasa Arab*, Surabaya: Usana Offset Printing, 1992.

Djaali dan Mulyono, *Pengukuran Dalam Bidang Pendidikan*, Jakarta, PT. Gramedia, 2008.

Fadliyanur, Pengelolaan Pengajaran Bahasa Arab,*Jornal Pendidikan,* http://fadliyanur.multiply.com/journalpendidikan/item/37.

Fayyad, Sulaiman, *An-Nahwu Al-Ashri, Dalil Mubsithliqawaid al-lughoh al-arabiyah*,Kairo: Al-Ahram, 1995.

Gronlund, Nrman E., *Measurement and Evaluation in Teaching.* New York: Machmillan Publishing Company, 1985.

Gunawan, Muhammad Ali, *Teknis Penyusunan Soal (Tes Dan Non Tes)*, Artikel, http://forumpenelitian.blogspot.com/2009/08/evaluasi-pendidikan-bagian-2.htm.

Hamalik, Oemar, *Evaluasi Kurikulum*, Bandung: Remaja Rosdakarya, 1993.

Hidayat, Yayat, *Studi Prinsip Dasar Metode Pengajaran Bahasa Arab,Artikel*, pp. 4-6 Tanggal 25 Desember 2009. http://arabicforall.or.id/metode/studi-prinsip-dasar-metode-pengajaran-bahasa-arab/

Hopkins, Kenneth D., Julian C. Stanley and R.R. Hopkins, *Education and Psychological Measurement,*Allyn and Bacon, 1990.

Hopkins, Charles D and Richard L. Antes, *Classroom Measurement and Evaluation,*Illionois, : F. E. Peacok Publisher, Inc., 1990.

Kerlinger, F.N, *Foundation OfBehavioural Research*, New York: Hill, Rinehart and Winston, 1973.

Lawshe.C.H, *A Quantitatif Approach to Content Validity*, Personel Psichology, 1975.

Livingston, Ronald B., *et. al., Measurement and Assesment in Education* Boston: Meril is an Inprint of Pearson, 2009.

McMillan, James H., *Assesment Essential for Standards-Based Education*, Cafornia: Carwin Press, A Sage Company, 2008.

Naga, Dali S., *Teori Pengukuran*, Jakarta: Program Pasca Sarjana Universitas Negeri Jakarta, 2008

_____, *Probibilitas dan Skor Pada Hipotesis Statistika*, Jakarta: UPT Taruma Negara, 2008.

Nitko, Anthony J., *Educational Assesmentof Student*, New Jersey: Simon &Scuster Company, Prentice Hall, Inc., 2001.

Peraturan Menteri Pendidikan Nasional, *Tentang Standar penilaian*, nomor 20 Tahun 2007.

Popham, W. James, *Modern Educational Measurement*, Los angeles: University of California, 1981.

_____, *Classroom Assesment, What Teacher Need to Know,* Needham Heights: Allyn Balon, A. Simon, & Schuster company, 1995.

Purwanto, *Evaluasi Hasil Belajar*, Yogyakarta: Pustaka Pelajar, 2009.

Purwanto, Ngalim, *Prinsip-prinsip dan Teknik Evaluasi Pengajaran*, Bandung: PT. Remaja Rosdakarya. 2008.

Sappaile, Pallawagau, *Pengaruh Tipe Soal Dan Waktu Testing Terhadap Daya Diskriminator Butir Soal Biologi* SMU, Tesis Jakarta: PPS UNJ, 2005.

Silverius, Suke*Evaluasi Hasil Belajar dan Umpan Balik*, Jakarta: Grasindo, 1991.

Sukardi, *Evaluasi Pendidikan Prinsip dan Operasionalnya,* Jakarta: Bumi Aksara, 2009.

Sudjana, Nana, *Penilaian Hasil Proses Belajar Mengajar***,** Bandung: PT. Remaja Rosdakarya, 2009.

_____, *Dasar-dasar Proses Belajar Mengajar*, Bandung: Sinar Baru Algesindo, 2005.

Sudijono, Anas, *Penganr Evaluasi Pendidikan*, Jakarta: Raja Grafindo Persada, 2005.

Sugiyono, *Metodologi Penelilian Kuantitatif, Kualitatif* dan R & G, Bandung; Alavabeta, 2008.

_____, *Statistik Untuk Penelitian*, Bandung: Alvabeta, 2007.

Surapratama, Sumarna, *Panduan Penulisan Tes Tertulis, Implementasi Kurikulum 2004*, Bandung: Remaja Rosdakarya, 2005.

_____, Analisis Validitas, Reliabilitas dan Interpretasi Hasil Tes, Jakarta:Remaja Rosdakarya, 2004.

Suryabrata, Sumadi, *Psikologi pendidikan*, Yogyakarta: Raja Grasindo Persada, 1993.

_____, *Pengembangan Tes Hasil Belajar* (Jakarta: Rajawali Press, 1987.

Sukadji, Soetarlinah, Validitas dan Reliabilitas*, Jurnal Pendidikan*, http://lussysf.multiply.com/journal pendidikan/item/137.

Syaiffudin*,* Azwar,Validitas dan Reliabilitas*, Jurnal Pendidikan,* http://lussysf.multiply.com/journal Pendidikan/item/137.

Undang-unm ubdang Tentang Sistem Pndidikan Nasional Nomor 20 Tahun 2003.

Uno, Hamzah B. , Herminanto Sofyan, dan I Made Candiasa, *Pengembangan Instrumen Untuk Penelitian*, Jakarta: Delima Press, 2001.

Widoyoko, Eko Putro, Evaluasi Program Pembelajaran, Yokyakarta: Pustaka Pelajar, 2009.

Wiersma, William, dan Stephen G. Jurs, *Educational Measurement and Testing*, Massachusetts: A division of Simon & Schuster, inc, 2001.