PENGEMBANGAN TES BERBASIS KOMPUTER

SRI MULIANAH
Universitas Negeri |Jakarta

WAHYU HIDAYAT
Universitas Kebangsaan Malaysia

ABSTRACT

This study aims to find practical procedure for applying the model to the test on the computer based test methodology of research subjects in order to develop tests in STAIN Parepare.This research method is a method of research and development, namely the development of computer-based test in the of research methodology subject . There are three stages in this study, namely Assembling Problem In Computer Systems, Calibration Test,Utilization In a limited scale .The results of research studies indicate that the development of the test can be done using a computer (computer base test) . The use of computers as a substitute for tests that use paper and pencil is more efficient and effective.

Keywords: *Computer Based Test and Examine.*

ABSTRAK

Penelitian ini bertujuan untuk menemukan prosedur praktis untuk menerapkan model untuk tes pada komputer berbasis metodologi tes subjek penelitian untuk mengembangkan tes di STAIN Parepare. Metode penelitian ini adalah metode penelitian dan pengembangan, yaitu pengembangan tes berbasis komputer dalam metodologi penelitian subjek. Ada tiga tahap dalam penelitian ini, yaitu Perakitan Masalah dalam Sistem Komputer, Kalibrasi Test dan Pemanfaatan dalam skala terbatas. Hasil studi penelitian menunjukkan bahwa pengembangan tes dapat dilakukan dengan menggunakan komputer (tes dasar komputer). Penggunaan komputer sebagai pengganti untuk tes yang menggunakan kertas dan pensil lebih efisien dan efektif.

Kata Kunci: *Tes Berbasis Komputer dan Periksa*

PENDAHULUAN

Perkembangan teknologi informasi telah menyentuh hampir semua sektor. Tak terkecuali sektor pendidikan. Dalam bidang pendidikan, teknologi informasi telah dimanfaatkan untuk menunjang layanan administrasi, proses pembelajaran, pendaftaran ulang, perpustakaan, akses nilai, pencarian referensi secara cepat dan mudah, proses penelitian, pembayaran SPP, bahkan untuk seleksi penerimaan mahasiswa baru. Keberadaan teknologi telah membantu sektor pendidikan menjadi lebih mudah, efektif dan efisien.

Di negara-negara maju penerapan teknologi sudah berlangsung lama. Dan penerapan teknologi informasi dalam proses pembelajaran telah mengubah model dan pola pembelajaran pada dunia pendidikan mereka. Ada banyak sistem pembelajaran yang menggunakan alat bantu komputer, salah satunya yaitu aplikasi pembelajaran yang mengacu pada teknologi berbasis Multimedia dan berbasis Web (Internet). *Computer-Based Instruction* (CBI) merupakan bentuk aplikasi komputer yang diterapkan dalam proses pembelajaran.

Pada awalnya, penerapan *Computer-Based Education popular* menggunakan program *Computer-Assisted Instruction* (CAI), *Computer-Assisted Learning* (CAL), *Computer-Managed Instruction* (CMI), dan *Computer-Assisted Guidance*. Begitupun

dalam sistem evaluasi pembelajaran khususnya sistem pengujian (*testing*) dapat juga memanfaatkan teknologi informasi, yaitu dilakukannya

Computer Based Test (CBT) atau evaluasi/ tes berbasis komputer. Peserta didik dapat melakukan tes dari tempat yang berbeda, baik itu dalam jaringan internet maupun dalam jaringan intranet dalam suatu organisasi. *Computer Based Test* dapat dijadikan sebagai sarana dalam evalusi pembelajaran. Dibebarapa sekolah evaluasi pembelajaran baik ulangan harian atau ujian sekolah masih menggunakan cara manual yaitu dengan *paper and pencil*. Cara ini dianggap tidak efisien dan praktis, diantaranya dalam hal biaya penyediaan bahan soal dan pemeriksaan.

Dengan model evaluasi pembelajaran memanfaatkan teknologi informasi sistem evaluasi pembelajaran akan lebih efektif dan efisien serta mampu melakukan evaluasi secara cepat, tepat dan memudahkan dalam melakukan pengukuran serta penilaian itu sendiri. Diharapkan semua kendala yang ditemui pada saat menjalankan cara manual dapat diperkecil atau bahkan dihilangkan.

Keunggulan dengan menggunakan aplikasi model CBT anatara lain : (1)Hasil tes dapat diketahui saat itu juga dengan cepat sesaat setelah peserta selesai mengikuti tes (hemat waktu). (2) Tidak perlu tim khusus untuk mengoreksi soal karena sistem yang akan langsung mengoreksi dan mengkalkulasi jumlah soal yang benar dan salah (hemat tenaga). (3) Tidak perlu menggandakan kertas-kertas soal dan lembar jawaban untuk dibagikan ke peserta tes (hemat biaya).(4) Dapat membangun bank soal

Perkembangan tes berbasis komputer akhir-akhir ini menjadi tren dibeberapa instansi termasuk instansi pemerintah. Salah satunya dalam penerimaan calon pegawai negeri sipili (CPNS). Pemerintah melaksanakan seleksi atau tes dengan memanfaatakan media komputer, yang disebut dengan computer asessment test (CAT). Ke depan sistem CAT akan diterapakan dalam penerimaan CPNS, karena sistem dianggap lebih efisien dan praktis.

Sekolah Tinggi Agama Islam Negeri (STAIN) Parepare sebagai perguruan tinggi yang sudah memiliki jaringan internet, sudah selayaknya untuk melakukan inovasi dalam mengembangkan tes berbasis komputer. selain pengembangan kelimuan dan teknologi, juga perlu dilihat kefektifan dan keefisien pengembangan tes seperti ini. Manfaat lain dari pengembangan tes ini adalah membangun bank soal yang terstandar. Hal ini amat jarang dilakukan di perguruan tinggi. Karena pengembangan *computer based test* (CBT) belum pernah dilakukan di STAIN Parepare. Maka sebagai ujicoba pengembangan *computer based test* (CBT), peneliti mengembangkannya pada mata kuliah *methodology of research*, dimana penulis mengampu 3 kelas di program pendidikan bahasa inggris.

Berdasarkan paparan di atas, penulis memandang perlunya sebuah pengembangan model tes dengan *Computer Based Test* (CBT). Alasan penulis mengkaji ini adalah : 1) Sebagai upaya mencari terobosan baru mengenai sistem evaluasi pembalajarn yang lebih efisien dan efektif, 2) Sistem tes di STAIN Parepare masih menggunakan *paper and pencil*.

Istilah tes bukanlah suatu istilah yang asing ditelinga kita. Tentunya makna tes yang dimaksud dalam penelitian ini adalah tes yang relevan dengan pengukuran (*measurement*) suatu prestasi belajar (*achievement learning*).

Menurut Linn & Gronlund dalam Wahyu Hidayat tentang tes

> "*an Instrument or systematic procedure for measuring a sample behaviour"* (Wahyu,2012:18).

Lee J. Cronbach menambahkan

> "*a systematic procedure for observing a person's behaviour and describing it with the aid of a numerical scale or a category system"*

Saifuddin Azwar, menarik kesimpulan tentang pengertian tes, antara lain:Tes adalah prosedur yang sistematis. Maksudnya item-item dalam tes disusun menurut cara dan

aturan tertentu; prosedur administrasi tes dan pemberian angka (*scoring*) terhadap hasilnya harus jelas dan dispesifikasikan secara terperinci; setiap orang yang mengambil tes itu harus mendapat item-item yang sama dalam kondisi sebanding.

Tes berisi sampel perilaku. Artinya betapun panjangnya suatu tes, item yang ada di dalamnya tidak akan dapat mencakup seluruh isi materi yang mungkin ditanyakan, dan kelayakan suatu tes tergantung pada sejauhmana item-item dalam tes itu mewakili secara representative kawasan (*domain*) perilaku yang diukur.

Tes mengukur perilaku. Artinya item-item dalam tes menghendaki agar subjek menunjukkan apa yang diketahui atau apa yang telah dipelajari subjek dengan cara menjawab pertanyaan-pertanyaan atau mengerjakan tugas-tugas yang dikehendaki tes.

Tes dilihat dari segi kegunaan untuk mengukur peserta didik, Suharsimi Arikunto membedakan atas adanya 3 macam tes menurut (azwar,2003:23) Tes diagnostik; Tes ini merupakan tes yang diberikan sesudah satu pelajaran disajikan, tujuannya adalah untuk mengetahui apakah peserta didik mendapat kesukaran pada bacaan tertentu dari pelajaran yang diberikan. Penyusunan tes untuk keperluan ini biasanya dititikberatkan pada materi dimana peserta didik melakukan banyak kesalahan atau banyak yang tidak bisa menjawab. Tes diagnostik bersangkut paut dengan usaha membentuk siswa mengatasi kesulitan belajarnya. Tes formatif; Tes ini merupakan tes yang diberikan sesudah satu kegiatan belajar mengumpulkan informasi tentang kekuatan dan kelemahan seseorang dalam pelajaran tersebut. Berkaitan dengan umpan balik yang dimaksudkan untuk acuan memperbaiki proses belajar mengajar.

Tes sumatif adalah adalah tes yang diberikan sesudah jumlah kegiatan belajar diselesaikan dalam satu periode tertentu tujuannya adalah untuk mengumpulkan data/ informasi dalam menentukan target dan taraf serap mahasiswa terhadap pelajaran yang telah diberikan. Berkaitan dengan pemberian nilai yang penentuan keputusan mengenai hasil atau kemampuan belajar siswa atau tes penempatan.

Selain itu, tes juga dapat diklasifikasikan berdasarkan waktu, yaitu (a) Kecepatan Test (*Speed Test*) Speed tes adalah tes yang didasarkan atau ditentukan oleh batasan waktu. Peserta tes dibatasi waktunya dalam mengerjakan soal tes. Ciri–cirinya yaitu waktu dibatasi dan tidak ada tingkat kesulitan. Contoh dari speed test adalah tes Skolastik atau tes potensi/kemampuan akademik. (b) Kekuatan Tes (*Power Test)Power Tes* adalah tes yang didasarkan sejauhmana kemampuan peserta tes mengerjakan soal tes. Peserta tes tidak dibatasi waktunya dalam mengerjakan soal tes. Contohnya tes kognitif

Computer Based Test (CBT);Tes biasanya dihubungkan dengan cara pengukuran terhadap penguasaan materi tertentu. Hasil dari tes salah satunya digunakan untuk membuat keputusan sekolah atau guru terhadap muridnya. Hasil tes dianggap sebagai bukti yang valid dari individu ,yang dapat digunakan misalnya untuk kenaikan kelas, promosi jabatan, dan kelulusan. Sebelum adanya tes berbasis komputer, biasanya tes dilakukan secara tertulis dalam kertas (*paper based test*), tetapi seiring dengan perkembangan teknologi informasi tes tertulis mulai bergeser digantikan dengan tes berbasis komputer bahkan internet.

Tes berbasis komputer (CBT) adalah metode penyajian tes sedemikian hingga respons peserta tes terhadap tes tersebut dapat disimpan dan dianalisis secara elektronik. Dengan kata lain tes berbasis komputer dilaksanakan dengan menggunakan bantuan *software* komputer).

Ada empat bentuk model tes berbasis komputer dan internet yang dikembangkan oleh ITC, yaitu (a) Terbuka (*Open Mode*); Tes dengan model terbuka seperti ini, dapat diikuti siapapun dan tanpa pengawasan siapapun,

contohnya tes yang dapat diakses secara terbuka di internet. Peserta tes tidak perlu melakukan registrasi peserta. (b)Terkontrol (*Controlled Mode*);Tes dengan model seperti ini, sama dengan tes dengan model terbuka yaitu tanpa pengawasan siapapun, tetapi peserta tes hanya yang sudah terdaftar, dengan cara memasukkan username dan password. (c) *Supervised Mode*; Pada model ini terdapat supervisor yang mengidentifikasi peserta tes untuk diotentikasi dan memvalidasi kondisi pengambilan tes. Untuk tes di internet mode ini menuntut administrator tes untuk meloginkan peserta dan mengkonfirmasi bahwa tes telah diselesaikan dengan benar pada akhir tes. (d) *Managed Mode*;Pada model ini biasanya tes dilaksanakan secara terpusat. Organisasi yang mengatur proses tes dapat mendefinisikan dan meyakinkan unjuk kerja dan spesifikasi peralatan di pusat tes.

Ada banyak keuntungan melakukan tes melalui komputer, diantaranya : mengijinkan melakukan tes di saat yang tepat bagi peserta, mengurangi waktu untuk pekerjaan penilaian tes dan membuat laporan tertulis, menghilangkan pekerjaan logistik seperti mendistribusikan, menyimpan dan tes menggunakan kertas.

Menurut Bjorner, Kosinski, dan Ware dan Bjorner bahwa kombinasi CBT maupun CAT dengan teori tes terutama TRB yang memanfaatkan bank soal dapat memberikan beberapa keuntungan antara lain, bank soal dapat diperluas secara berangsur-angsur dengan menambahkan soal ataupun mengevaluasi butir soal yang ada, dan proses respons peserta dapat dipantau/dimonitor untuk memastikan mutu penilaian dan pola respons yang tidak konsisten dapat diselidiki.

Pada dasarnya pelaksanaan *Computer Based Test* sama halnya dengan proses pembelajaran menggunakan komputer. *Computer Based Test* atau tes berbasis komputer dapat dilaksankan dalam laboratorium komputer yang telah terkoneksi dengan jaringan dan sistemnya. Dalam pelaksanaan tes berbasis komputer (CBT) ada beberapa hal yang perlu diperhatikan diantaranya : ke-ontetikan peserta test, bank soal, sistem Computer-based test itu sendiri.

Proses otentikasi dalam tes berbasis komputer (CBT), merupakan hal yang sangat penting, untuk menentukan siapa saja yang bisa mengikuti tes. Biasanya dalam proses ini, peserta tes akan diberikan sebuah username dan password, yang akan digunakan untuk login sehingga peserta dapat masuk dan mengikuti tes.

Ketersediaan soal dalam jumlah yang cukup banyak menjadi syarat selanjutnya dalam tes berbasis komputer (CBT). Dari jumlah soal yang cukup banyak memungkinkan pemilihan soal secara random sehingga antar peserta tes akan mendapatkan soal yang berbeda. Hal ini dilakukan untuk menghindari adanya kerjasama antara peserta test.

Sistem *Computer Based Test* yang telah melalui uji kelayakan sangat diperlukan, mengingat pada umumnya tes berbasis komputer dilaksanakan dalam waktu yang sama. Sehingga dibutuhkan software dan hardware yang mendukung, istilah dalam teknologi informasi yaitu client-server. Di mana komputer peserta tes (client) terhubung dengan sistem tes berbasis komputer melalui komputer server. Dalam hal ini jumlah client jauh lebih banyak dari jumlah server, untuk itulah dibutuhkan sistem tes berbasis komputer yang layak pakai.

Pelaksanaan pengukuran di bidang pendidikan pada prinsipnya bertujuan untuk mengetahui karakteristik suatu objek seperti kemampuan, keberhasilan belajar, sikap, minat atau ciri terpendam lainnya yang terdapat pada peserta didik namun tidak kelihatan dan tidak dapat diukur langsung. Untuk mengukur berbagai karakteristik yang terpendam itu sangat diperlukan alat ukur yang baik sehingga mampu mengungkap secara benar ciri terpendam pada peserta didik. Alat ukur yang baik adalah alat ukur yang memenuhi persyaratan dan mampu menghasilkan

informasi yang mengandung kesalahan sekecil mungkin.

Parameter yang digunakan pada analisis butir berdasarkan teori tes klasik dan teori respons butir pada dasarnya adalah sama yaitu tingkat kesukaran, daya pembeda, tebakan semu *(pseudo guessing),* dan kemampuan. Perbedaanya terletak pada formula, skala, dan satuan yang digunakan. Selain itu, analisis butir suatu tes dengan teori tes klasik dan teori respons butir pada prinsipnya juga dilakukan untuk menaksir kemampuan seseorang yang diharapkan memiliki kesalahan sekecil mungkin. Kesalahan pengukuran menurut teori tes klasik dinyatakan dengan kesalahan baku pengukuran (*Standar Error of Measurement/ SEM*) yang besarnya tergantung pada indeks kehandalan tes. Untuk teori respons butir kesalahan pengukuran dinyatakan dengan kesalahan baku pengukuran (*StandarError of Measurement/SEM*) yang besarnya tergantung pada tingkat kemampuan seseorang dan fungsi informasi tes. Adanya kesalahan yang melekat pada data hasil pengukuran ini disebabkan oleh banyak faktor diantaranya adalah alat ukur itu sendiri, pelaksanan pengukuran, objek pengukuran, dan teknik analisis yang digunakan.

Teori Tes Klasik

Analisis butir soal secara klasik adalah proses penelaahan butir soal melalui informasi dari jawaban peserta didik guna meningkatkan mutu butir soal yang bersangkutan dengan menggunakan teori tes klasik. Kelebihan analisis butir soal secara klasik adalah murah, dapat dilaksanakan seharihari dengan cepat menggunakan komputer, murah, sederhana, familier dan dapat menggunakan data dari beberapa peserta didik atau sampel kecil. Aspek yang perlu diperhatikan dalam analisis butir soal secara klasik adalah setiap butir soal ditelaah dari segi: tingkat kesukaran butir, daya pembeda butir, dan penyebaran pilihan jawaban (untuk soal bentuk obyektif) atau frekuensi jawaban pada setiap pilihan jawaban.

Tingkat Kesukaran Butir (item difficulty index)

Tingkat kesukaran butir soal adalah peluang menjawab benar suatu soal pada kemampuan tertentu yang biasanya dinyatakan dalam bentuk indeks. Besarnya indeks kesukaran antara 0,00 sampai dengan 1,00 (Aiken : 66) (Linn,2000:55).

0,00 ⟶ 1,00

Semakin besar indeks tingkat kesukaran maka semakin mudah soal itu. Berdasarkan indeks tingkat kesukaran maka seharusnya lebih tepat jika disebut tingkat kemudahan butir soal. Karena sudah menjadi kesepakatan para ahli maka sampai sekarang masih tetap menggunakan istilah tingkat kesukaran butir soal. Makna tingkat kesukaran (TK) = 1, artinya bahwa peserta tes menjawab benar soal itu, TK = 0, artinya tidak ada peserta tes yang menjawab benar pada soal.

Pada prinsipnya taraf kesukaran tes bentuk soal ini dihitung berdasarkan proporsi jumlah peserta yang menjawab benar terhadap jumlah total peserta tes.

$$P = \frac{B}{T}$$

Keterangan :

P = taraf sukar butir

B = jumlah peserta yang menjawab benar (*item score*)

T = jumlah total peserta tes

Tingkat Kesukaran tes atau dapat dihitung berdasarkan jumlah peserta yang menjawab benar pada kelompok atas dan kelompok bawah yang dirumuskan sebagai berikut :

$$TK = \frac{BA + BB}{N}$$

Keterangan :

TK = Tingkat Kesukaran

BA = jumlah jawaban benar pada kelompok atas (27 %)

BB = jumlah jawaban benar pada kelompok bawah (27 %)

N = ukuran kelompok (jumlah peserta kelompok atas dan bawah)

Kriteria Indeks Kesulitan Butir Soal:

- 0,00 - 0,30 = Soal kategori sukar
- 0,31 - 0,70 = Soal kategori sedang
- 0,71 - 1,00 = Soal kategori mudah

Daya pembeda butir soal tes mengacu pada kemampuan butir dalam membedakan kemampuan antara peserta tes yang telah menguasai materi dan peserta tes yang tidak/ belum menguasai materi yang ditanyakan. Daya pembeda dinyatakan dalam indeks. Indeks daya pembeda berkisar antara -1,00 sampai dengan +1,00. Semakin tinggi indeks daya pembeda soal artinya semakin semakin mampu soal yang bersangkutan membedakan peserta tes /peserta tes yang telah memahami materi dengan peserta tes yang belum memahami materi. Semakin tinggi daya pembeda suatu butir soal, maka semakin kuat/baik butir soal tersebut. Jika indeks daya pembeda bernilai negatif (DP < 0), berarti lebih banyak kelompok bawah (peserta tes / peserta tes yang belum memahami materi) menjawab benar soal tersebut dibandingkan dengan kelompok atas (peserta tes /peserta tes yang memahami materi).

Dalam seleksi item, setiap item yang memiliki indeks lebih besar dari 0,50 dapat langsung dianggap sebagai item yang berdaya diskriminasi baik, item yang memiliki indeks kurang dari 0,20 dapat langsung dibuang, sedangkan item lainnya dapat ditelaah lebih lanjut untuk direvisi (Crocker,1986:315).

Klasifikasi/Kriteria Daya Pembeda :

- 0,40 – 1,00 Soal diterima/baik
- 0,30 – 0,39 Soal diterima tetapi perlu perbaikan
- 0,20 – 0,29 Soal diperbaiki
- 0,19 – 0,00 Soal tidak dipakai/dibuang

Validitas berasal dari kata *validity* (shahih dalam bahasa arab) yang mempunyai arti sejauh mana ketepatan dan kecermatan suatu alat ukur dalam melakukan fungsi ukurnya.

Suatu alat tes dapat dikatakan mempunyai validitas yang tinggi apabila alat tes tersebut menjalankan fungsi ukurnya, atau memberikan hasil ukur yang sesuai dengan maksud dilakukannya pengukuran tersebut. Sedangkan tes yang memiliki validitas rendah akan menghasilkan data yang tidak relevan dengan tujuan pengukuran.

Validitas alat tespada umumnya digolongkan dalam tiga kategori, yaitu : (a) Validitas Konstruksi (*Construct Validity*) Validitas konstruk adalah validitas yang menyangkut bangunan teoretik variabel yang akan diukur. Sebuah tes dikatakan mempunyai validitas konstruk apabila butir-butir soal yang disusun dalam tes mengukur setiap aspek berpikir dari sebuah variabel yang akan diukur melalui tes tersebut. Untuk menguji validitas konstruksi, dapat digunakan pendapat dari para ahli (*Judgmen Expert*). Para ahli diminta pendapatnya tentang alat tes tersebut. (b) Validitas Isi (*Content Validity*);Validitas isi disebut juga validitas kurikuler. Oleh karena itu, validitas ini erat kaitannya dengan materi yang akan diukur dalam tes. Tentu saja materi yang dimaksud adalah materi yang terdapat dalam kurikulum. Pengujian validitas isi dapat dilakukan dengan membandingkan antara isi alat tes dengan isi atau rancangan yang telah ditetapkan. Validitas isi mencerminkan sejauh mana butir-butir dalam tes mencerminkan materi yang disajikan dalam kurikulum. Sebuah tes dikatakan memiliki validitas isi jika butir-butir tes bersifat representatif terhadap isi materi dalam kurikulum tersebut. Pengujian validitas isi tidak melalui prosedur pengujian secara statistik, melainkan melalui analisis secara rasional. Pengetahuan terhadap kurikulum menjadi dasar berpijak yang penting untuk dapat melakukan analisis validitas isi. Cara yang praktis untuk melakukan analisis validitas isi adalah dengan melihat apakah

butir-butir tes telah disusun sesuai dengan *blue-print* (kisi-kisi) yang sudah dirancang sebelumnya. *Blue print* menjadi acuan.

Sesuai dengan namanya, validitas ini didasarkan pada kriteria tertentu. Dengan demikian bukti adanya validitas ditunjukkan adanya hubungan korelasional skor pada tes yang bersangkutan dengan skor suatu kriteria.

Pengujian validitas ini bersifat empirik, artinya pengujian hanya dapat dilakukan setelah mendapatkan data di lapangan. Apabila berdasarkan hasil analisis yang dilakukan terhadap data hasil pengamatan di lapangan terbukti bahwa tes hasil belajar dapat mengukur hasil belajar yang seharusnya diungkap secara tepat maka berarti alat tes tersebut mempunyai validitas empirik. Untuk keperluan pengujian jenis validitas ini dapat dilakukan dengan dua cara yaitu dari segi kemampuannya dalam melakukan ramalan (*predictive validity*) serta daya ketepatan bandingannya (*concurent validity*).

Perbedaan utama antara validitas ramalan dengan validitas bandingan adalah ketersediaan pembanding (kriterium). Pada validitas ramalan, kriterium diperoleh pada waktu yang akan datang setelah dilakukan tes yang akan diukur validitasnya tersebut. Sedangkan pada validitas bandingan, kriterium sudah ada atau dapat diperoleh pada saat yang sama dengan waktu untuk memperoleh data tentang tes yang akan diukur validitasnya tersebut tanpa harus menunggu masa yang akan datang. Uji validitas butis soal pilihan ganda menggunakan korelasi point biserial yaitu korelasi antara data interval dan data dikotomi.

$$r_{pbis} = \frac{\bar{X}_b - \bar{X}_s}{SD_t}\sqrt{pq}$$

Keterangan :

$\bar{X}_b$ = rata-rata skor siswa/peserta tes yang menjawab benar

$\bar{X}_s$ = rata-rata skor siswa/peserta tes yang menjawab salah

SD_t = simpangan baku skor total

p = proporsi jawaban benar terhadap semua jawaban siswa/peserta tes

q = 1 - p

Selain validitas, alat ukur yang baik juga harus reliabel. Oleh karena itu, alat ukur yang baik adalah alat ukur yang valid dan reliabel. Dalam kajian teoritis, reliabilitas adalah sejauh mana pengukuran dari suatu uji coba yang dilakukan tetap memiliki hasil yang sama meskipun dilakukan secara berulang-ulang terhadap subjek dan dalam kondisi yang sama. Instrumen alat ukur dianggap bisa diandalkan apabila memberikan hasil yang konsisten untuk pengukuran yang sama dan tidak bisa diandalkan bila pengukuran yang dilakukan secara berulang-ulang itu memberikan hasil yang relatif tidak sama. Pengujian reliabilitas instrumen untuk memperoleh hasil yang reliabel bisa dilakukan dengan berbagai metode statistik.

Ada 3 cara yang dapat dilakukan untuk menentukan reliabilitas skor tes, yaitu : (a) Metode Tes Ulang (*Test Retest Method*) diterapkan untuk menghindari adanya penyusunan dua seri tes. Teknisnya adalah sebuah tes yang sama diberikan dua kali kepada responden yang sama dengan jarak waktu tertentu. Jika hasil tes pertama mempunyai kesejajaran dengan hasil tes yang kedua maka tes tersbut dikatakan reliable. Oleh karena pengujian ini dilakukan terhadap sebuah tes yang diujicobakan dua kali maka sering disebut pula sebagai *single-test-double-trial-method*. Kelemahan metode ini adalah jika jeda waktu tes terlalu singkat sedangkan soal tes banyak mengungkapkan aspek pengetahuan maka responden cenderung masih mengingat materi yang diteskan, sehingga ada kemungkinan hasil tes yang kedua lebih baik daripada hasil tes pertama. Sebaliknya jika jeda waktu tes pertama dengan

kedua terlalu lama dikhawatirkan banyak faktor serta situasi dan kondisi sudah banyak berubah dan mempengaruhi hasil tes yang kedua. (b) Metode Tes Sejajar (*Equivalent*) mengharuskan adanya dua buah seri soal yang mempunyai kesamaan tujuan, bobot soal, tingkat kesukaran, susunan soal, tetpai butir –butir soalnya berbeda. Dengan kata lain, dua buah tes yang digunakan harus sejajar (paralel, equivalen). Koefisien relibiabilitas diperoleh dengan mengkorelasikan hasil tes pertama dengan hasil tes kedua. Sudah tentu metode ini akan menambah kerepotan. Inilah kelemahan metode ini. Kelebihan dari metode ini adalah dapat memperbaiki kelemahan pada metode pertama yaitu terhindarnya dari kondisi "siswa masih mengingat materi tes pertama". Aspek ingatan dan hafalan pada pengerjaan tes pertama tidak terbawa pada saat mengerjakan tes yang kedua.(c) Metode Belah Dua (Split – Half) ini dari kepraktisannya lebih praktis dari pada dua metode sebelumnya. Metode ini hanya melakukan sekali tes kepada sekelompok subjek. Dengan demikian tidak perlu menunggu waktu maupun harus mempunyai data dari tes sejenis untuk dapat menentukan reliabilitasnya. Reliabilitas diukur hasil pengukuran belahan pertama dan belahan kedua dari alat ukur yang sama.

Untuk menentukan reliabilitas alat ukur maka digunakan kriteria Kaplan (Kaplan & Sacuzzo, 2005), yaitu:

R ≥ 0,70 = alat ukur dapat diandalkan (kurang reliabel)

R < 0,70 = alat ukur kurang dapat diandalkan (reliabel)

Untuk mengetahui koefisien reliabilitas tes soal bentuk pilihan ganda digunakan rumus Kuder Richadson 20 (KR-20) seperti berikut ini.

Rumus Kuder Richardson 20 (KR – 20).

$$KR-20=\frac{k}{k\text{-}1}\left(1-\frac{\sum pq}{SD_t^2}\right)$$

Keterangan

k = jumlah butir soal

SD_t^2 = varian skor total

P = proporsi siswa yang menjawab benar

q = 1 -p

Untuk analisis butir soal bisa digunakan analisis butir secara modern yaitu dengan penelaahan butir soal dengan menggunakan *Item Response Theory* (IRT) atau teori jawaban peserta tes. Teori ini meruapakan teori yang menggunakan fungsi matematika untuk menghubungkan antara peluang jawaban benar suatu soal dengan kemampuan peserta tes. Nama lain dari IRT adalah *Latent Trait Theory* (LTT) atau *Characteristic Curve Theory* (CCT).

Untuk mengetahui kelebihan analisis IRT, maka para evaluator perlu mengetahui keterbatasan analisis secara klasik. Keterbatasan model pengukuran secara klasik bila dibandingkan dengan teori jawaban butir soal adalah seperti berikut (Hambleton,1991:25) (1) Tingkat kemampuan dalam teori klasik adalah "true score". Jika tes sulit artinya tingkat kemampuan peserta didik mudah. Jika tes mudah artinya tingkat kemampuan peserta didik tinggi. (2) Tingkat kesukaran soal didefinisikan sebagai proporsi peserta didik dalam grup yang menjawab benar soal. Mudah/sulitnya butir soal tergantung pada kemampuan peserta didik yang dites dan kemampuan tes yang diberikan. (3) Daya pembeda, reliabilitas, dan validitas soal/tes didefinisikan berdasarkan grup peserta didik.

Asal mula IRT adalah kombinasi suatu versi hukum phi-gamma dengan suatu analisis faktor butir soal (*item factor analysis*) kemudian bernama Teori Trait Laten (*Laten Trait Theory*), kemudian sekarang secara umum dikenal IRT.

Munculnya IRT didasari dari kelemahan analisis secara klasik, yaitu : Abilitas dalam teori klasik adalah *true score*. Artinya jika tes sulit artinya abilitas peserta tes rendah. Dan jika tes mudah artinya abilitas tinggi. Mudah/ sulitnya butir soal tergantung pada kemampuan

peserta tes yang dites dan kemampuan tes yang diberikan. Daya pembeda, reliabilitas, dan validitas tes/soal didefinisikan berdasarkan kelompok peserta tes. Sedangkan kelebihan IRT adalah : IRT tidak berdasarkan kelompok dependent, Skor peserta tes dideskripsikan bukan tes dependent, Model ini menekankan pada tingkat butir soal bukan tes, IRT tidak memerlukan pararel tes untuk menentukan reliabilitas tes dan IRT suatu model yang memberikan suatu pengukuran ketepatan untuk setiap skor reliabilitas.

Tujuan utama IRT adalah memberikan kesamaan antara statistik soal dan estimasi kemampuan. Ada tiga keuntungan IRT, yaitu : (1) asumsi pada populasi tingkat kesukaran, daya pembeda merupakan independen, (2) asumsi pada populasi tingkat kesukaran, daya pembeda merupakan independen sampel yang menggambarkan untuk tujuan kalibrasi soal, (3) statistik yang dipergunakan untuk menghitung tingkat kemampuan peserta tes diperkirakan dapat terlaksana.

Ada empat macam model IRT, yaitu (1) Model satu parameter (model Rasch), yaitu untuk menganalisis data yang hanya menitikberatkan pada parameter tingkat kesukaran soal. (2) Model dua parameter, yaitu untuk menganalisis data yang hanya menitikberatkan pada parameter tingkat kesukaran dan daya pembeda soal. (3)Model tiga parameter, yaitu untuk menganalisis data yang menitikberatkan pada parameter tingkat kesukaran soal, daya pembeda soal dan menebak. (4)Model empat parameter, yaitu untuk menganalisis data yang menitikberatkan pada parameter tingkat kesukaran soal, daya beda soal, menebak dan penyebab lain.

Easy Quiz merupakan perangkat lunak untuk pembuatan soal, kuis atau tes secara online (berbasis web). Penggunaan *Easy Quiz* dalam pembuatan soal tersebut sangat familiar/*user friendly*, sehingga sangat mudah digunakan dan tidak memerlukan kemampuan bahasa pemrograman yang sulit untuk mengoperasikannya.

Hasil soal, kuis dan tes dibuat/disusun dengan perangkat lunak ini dapat disimpan dalam format Flash yang dapat berdiri sendiri (*stand alone*) di website. Dengan *Easy Quiz*, pengguna dapat membuat dan menyusun berbagai bentuk dan level soal yang berbeda, yaitu bentuk soal benar/salah (*true/false*), pilihan ganda (*multiple choices*), pengisian kata (*fill in the blank*), penjodohan (*matching*), Kuis dengan area gambar dan lain-lain. Bahkan dengan *Easy Quiz* dapat pula disisipkan berbagai gambar (*images*) maupun file Flash (*Flash movie*) untuk menunjang pemahaman peserta didik dalam pengerjaaan soal.

Beberapa fasilitas yang tersedia dalam *Easy Quiz* selain dari sisi kemudahan penggunaan (*user friendly*) soal-soal yang dihasilkan, diantaranya yaitu (1). Fasilitas umpan balik (*feed-back*) berdasar atas respon/jawaban dari peserta tes, (2). Fasilitas yang menampilkan hasil tes/score dan langkah-langkah yang akan diikuti peserta tes berdasar respon/ jawaban yang dimasukkan, (3). Fasilitas mengubah teks dan bahasa pada tombol dan label sesuai dengan keinginan pembuat soal, (4). Fasilitas memasukkan suara dan warna pada soal sesuai dengan keinginan pembuat soal, dan (5). Fasilitas hyperlink; yaitu mengirim hasil/score tes ke email atau LMS. (6) Fasilitas pembuatan soal random, (7) Fasilitas keamanan dengan User account/password, (8) Fasilitas pengaturan tampilan yang dapat di modifikasi, dll.

Kriteria yang digunakan peneliti untuk mengembangkan tes diagnostik berbasis komputer ini mengacu pada kriteria kualitas suatu material yang dikemukakan oleh Nieveen. Menurut Nieveen (1999) suatu material dikatakan berkualitas jika memenuhi aspek-aspek kualitas produk antara lain: kevalidan (*validity*), kepraktisan (*practicality*), dan keefektifan (*effectiveness*). (Kevalidan (*validity*) Menurut (Nieven,1999) aspek validitas dari material dilihat dari apakah berbagai komponen dari material itu terkait secara konsisten antara satu dengan

yang lainnya. Sedangkan Arikunto (2008) menjelaskan bahwa suatu tes dikatakan valid jika tes tersebut dapat mengukur apa yang hendak diukur dengan tepat. Validitas tes ditinjau dari berbagai segi yaitu: validitas materi, validitas konstruksi (isi), dan validitas bahasa. Berdasarkan definisi kevalidan dari para ahli, maka kriteria kevalidan tes yang dikembangkan pada penelitian ini meliputi: validitas materi yaitu kesesuaian soal dengan indikator yang telah ditentukan, validitas konstruksi yaitu sistematika penulisan soal dan pilihan jawaban, validitas bahasa yaitu penggunaan bahasa yang sesuai ejaan yang disempurnakan (EYD) pada penulisan soal. (Kepraktisan (*practicality*), (Menurut Nieveen,1999) aspek kepraktisan dari material dilihat dari kemudahan material dapat digunakan. Keefektifan (*effectiveness*). Maka model tes yang dikembangkan peneliti dikatakan efektif dilihat dari komponen-komponen antara lain: Kesesuaian hasil tes dengan tujuan tes serta respons dosen dan respons mahapeserta tes mengenai keefektifan tes.

ITEMAN merupakan program komputer yang digunakan untuk menganalisis butir soal secara klasik. Program ini termasuk satu paket program dalam MicroCAT yang dikembangkan oleh *Assessment Systems Corporation* dimulai tahun 1982 dan mengalami revisi pada tahun 1984, 1986, 1988, dan 1993; mulai dari versi 2.00 sampai dengan versi 3.50. Assessment Systems Corporation beralamat di 2233 University Avenue, Suite 400, St Paul, Minesota 55114, United States of America.

Pengembangan tes berbasis komputer (CBT) hakikatnya memindahkan tes yang biasanya menggunakan paper and pencil ke dalam sistem komputer dengan bantuan software yang ada. Untu lebih praktisnya, dapat digunakan software yang sudah ada seperti *software easy quiz*. Setelah tes dirakit dalam sistem komputer maka langkah selanjutnya melakukan kalibrasi atau standarisasi tes dengan melakukan ujicoba terbatas kepada beberapa mahasiswa. Tujuan dari ujicoba ini untu melihat kualitas butir soal seperti tingkat kesukaran, validitas dan reliabilitas. Analisis butir soal dalam penelitian ini menggunakan software ITEMAN.

Tes yang sudah dikalibrasi selanjutnya dapat digunakan dalam skala yang terbatas. Tes pada penelitian ini digunakan untuk mid semester pada mata kuliah *methodology of research* pada program pendidikan bahasa inggris semester 5 di STAIN Parepare. Tes ini menggunakan tiga paket soal, yaitu paket soal 1, paket soal 2 dan paket soal 3.

Jika mengacu pada kualitas dan bobot yang sama pada ketiga paket soal tersebut, maka diduga tidak ada perbedaan hasil tes ketiga kelompok peserta tes dengan ketiga paket soal tersebut.

Metode penelitian ini merupakan metode penelitian pengembangan dan Penelitian (*research and development*), yaitu pengembangan ujian berbasis komputer pada mata kuliah *methodology of research*. Ada tiga tahap dalam penelitian ini, yaitu : Perakitan Soal Pada Sistem Komputer, Kalibrasi Tes dan Pemanfaatan Pada Skala terbatas

Pengembangan Model dan Prosedur Pengembangan

Hasil yang diharapkan dari pengembangan model ini adalah suatu program *computer based test* (CBT) yang diimplementasikan pada ujian mata kuliah *methodology of research*. Model ini kemudian diujicobakan dan dikalibrasi dengan menggunakan analisis IRT.

Prosedur pengembangan dalam penelitian ini terdiri atas beberapa tahap, yaitu (1) Tahap identifikasi bidang standar kompetensi yang akan diujikan pada ujian mata kuliah *methodology of research*. Selain itu perlu ditentukan terlebih dahulu indikator-indikator pada tiap bidang kompetensi yang akan diujikan . (2) Tahap menyusun soal dan bank soal sesuai dengan indikator bidang kompetensi yang akan diujikan. Setelah penentuan indikator pada tiap bidang kompetensi yang akan diujikan,

kemudian dibuatkan kisi-kisi soal. Setiap indikator terdiri dari 2 soal dengan kualitas yang sama. Penyusunan soal diperlukan sebelum diaplikasikan ke dalam sistem komputer. (3) Tahap membuat program CBT Setelah dipersiapkan soal, maka langkah selanjutnya adalah pembuatan program CBT dengan menggunakan *software easy quiz*. (4) Tahap Implementasi CBT Tahap akhir dari pengembangan model ini adalah kalibrasi yang dilakukan pada peserta tes tingkat akhir yang akan mengikuti ujian. kalibrasi ini diperlukan untuk melihat kualitas soal, apakah soal tersebut sudah standar atau belum. Untuk analisis standar digunakan analisis klasik atau analisis modern (IRT).

Populasi pada penelitian ini adalah seluruh mahasiswa yang mengikuti perkuliahan *methodology of research* porgram pendidikan bahasa inggris STAIN Parepare Semester lima. Adapun sampling frame penelitian ini adalah 36 mahasiswa pada pemanfaatan skala terbatas, yaitu mid semester mata kuliah *methodology of research*. Teknik pengambilan sampel menggunakan multistage random.

Penelitian ini merupakan penelitian pengembangan untuk menghasilkan model ujian pada mata kuliah *methodology of research*. Data yang diperoleh dalam penelitian ini adalah yang terkait dengan : Data Hasil Kalibrasi dan Data pemanfaatan pada skala terbatas (mid semester mata kuliah *methodology of research*

Untuk data yang terkait dengan kalibrasi butir soal pada ujicoba intrumen. Data yang terekam (data dokumentasi), kemudian dianalisis dengan analisis modern, yaitu Item Response Theory (IRT). Untuk mempermudah analisis data menggunakan komputer. Analisis butir soal dengan komputer maksudnya adalah penelaahan butir soal secara kuantitatif yang penghitungannya menggunakan bantuan program komputer. Analisis data dengan menggunakan program komputer adalah sangat tepat. Karena tingkat keakuratan hitungan dengan menggunakan program komputer lebih tinggi bila dibandingkan dengan diolah secara manual atau menggunakan kalkulator/ tangan. Program komputer yang digunakan untuk menganalisis data modelnya bermacam-macam tergantung tujuan dan maksud analisis yang diperlukan.

Program yang sudah dikenal secara umum adalah EXCEL, SPSS (*Statitistical Program for Social Science*), atau program khusus seperti ITEMAN (analisis secara kiasik), RASCAL, ASCAL, BILOG (analisis secara item respon teori atau IRT), FACETS (analisis model Rasch untuk data kualitatif.

Untuk penelitian ini analisis data digunakan program ITEMAN.Tahap awal dalam mengoperasikan ITEMAN adalah membuat "file data" (control tile) yang berisi lima komponen utama. Baris pertama adalah baris pengontrol yang mendeskripsikan data, Baris kedua adalah daftar kunci jawaban setiap butir soal, Baris ketiga adalah daftar jumlah option untuk setiap butir soal,Baris keempat adalah daftar butir soal yang hendak dianalisis (jika butir yang akan dianalisis diberi tanda Y (yes), jika tidak diikutkan dalam analisis diberi tanda N (no) dan Baris kelima dan seterusnya adalah data siswa dan pilihan jawaban siswa.

Cara menggunakan program ini, pertama data diketik di DOS atau Windows. Cara termudah adalah menggunakan program Windows yaitu dengan mengetik data di tempat Notepad.

Sedangkan untuk data yang terkait dengan pemanfaatan skala terbatas, yaitu pada mid semester mata kuliah *methodology of research*. Analisisnya menggunakan uji perbedaan *one way anova* antara ketiga kelompok untuk masing-masing kelompok dengan paket soal yang berbeda tetapi dari segi kualitas dan bobot soal sama. Adapun hipotesis statistik penelitian pada ketiga kelompok tersebut adalah :

$H_0 : \mu_1 = \mu_2 = \mu_3$

H_1 : Ada salah satu tanda yang tidak sama

Untuk mempermudah dalam menganalisis data, penulis memanfaatkan software SPSS Versi 22.0.

PEMBAHASAN

Dalam pembuatan tes dengan perangkat lunak (software) bisa digunakan software yang sudah ada seperti *wondershare quiz creator, test creator* dan *easy creator*. Software tersebut memiliki kelebihan dan kekurangnnya masing-masing. Untuk penelitian ini peneliti menggunakan *software easy quiz*. Selain mudah untuk digunakan, hasil test program *software easy quiz* langsung dibackup dalam word. Sehingga hasilnya dengan mudah dapat dianalisis.

Langkah dalam pembuatan tes dengan *software esy quiz* adalah sebagai berikut:

Pembuatan materi soal yang akan diteskan disesuaikan dengan silabus mata kuliah *methodology or research* yang pada pertemuan pertama kuliah sudah diberitahukan kepada mahasiswa. Butir soal yang diberikan sebanyak 20 soal dengan paket soal sebanyak 3 paket. Paket soal ini setara dalam indikator dan kompetensi dasarnya, begitupun secara validitas konseptual terutama dalam kontennya setara antar paket soal. Penggunaan 3 paket soal untuk menghindari kebocoran soal.

Soal yang sudah dirakit dimasukkan dalam *software easy quiz* yang sudah tersedia.

Setelah menginput dan mensetting maka langkah selanjutnya adalah mengklik set run.

Kualitas tes, termasuk bentuk tes pilihan ganda (dikotomi) dapat diungkap melalui analisis butir soal secara teoretis (telaah) dan analisis empiris. Analisis butir soal secara kualitatif dilakukan untuk menilai butir soal ditinjau dari aspek materi,konstruksi, dan bahasa. Analisis secara kuantitatif menekankan pada analisiskarakteristik butir soal secara empiris. Karakteristik butir soal antara lain meliputi indeks kesukaran *(p)*, daya beda *(d)*, dan distribusi respons.

Analisis secara empiris dapat menggunakan pendekatan tes klasik (*Clasical Test Theory* atau CTT) maupun pendekatan tes modern (*Item Respons Theory* atau IRT). Pada penelitian ini, penulis menggunakan sofware ITEMAN untuk analisis butir soal. Berikut analisis butir soal untuk tiap-tiap paket soal :

a. Paket Soal 1

Untuk analisis butir soal digunakan tingkat kesukaran, Validitas dan reliabilitas butir soal dengan menggunakan software ITEMAN dalam analisis.

Berdasarkan hasil analisis IRT (ITEMAN), maka dapat disimpulkan tingkat kesukaran (*Difficulty Item Index*) sebagai berikut :

Tabel 1

Rekapitulasi Tingkat Kesukaran Butir Soal Paket Soal 1

Butir Soal	Keterangan
1	Sedang
2	Sedang
3	Sedang
4	Sedang
5	Sukar
6	Sedang
7	Sedang
8	Sedang
9	Sedang
10	Sedang
11	Sedang
12	Sedang
13	Sedang
14	Sukar
15	Sedang
16	Sedang
17	Sedang
18	Sedang
19	Mudah
20	Sedang

Tabel 1 di atas menujukkan bahwa tingkat kesukaran paket soal 1(85%) dalam kategori sedang.

Analisis kualitas butir paket soal 1 selanjutnya adalah analisis validitas atau kriteria baik tidaknya butir soal. Menurut Ebel dan Frisbie dalam *Essentials of EducationalMeasurement* Kriteria baik tidaknya butir soal adalah bila korelasi point biserial: >0.40 = butir soal sangat baik; 0.30 -0.39 = soal baik, tetapi perlu perbaikan; 0.20 - 0.29 = soal dengan beberapa catatan,

biasanyadiperlukan perbaikan; < 0. 19 = soal jelek, dibuang, atau diperbaiki melalui revisi. Berikut kesimpulan kriteria baik tidaknya butir soal pada paket soal 1 :

Tabel 2

Rekapitulasi Kriteria Baik Tidaknya Butir Soal Paket Soal 1

Butir Soal	Kualitas Butir
1	sangat baik
2	sangat baik
3	sangat baik
4	sangat baik
5	sangat baik
6	sangat baik
7	Dibuang
8	Dibuang
9	Dibuang
10	sangat baik
11	sangat baik
12	Dibuang
13	sangat baik
14	sangat baik
15	Dibuang
16	Baik
17	Baik
18	Baik
19	Baik
20	Baik

Berdasarkan kriteria Ebel dan Frisbie, maka butir soal pada paket soal 1 yang didrop atau tidak dipergunakan yaitu butir soal 7,8,9,12 dan 15. Sehingga hanya 15 butir soal yang memenuhi standar atau baku.

Selanjutnya analisis kalibrasi adalah reliabilitas tes secara keseluruhan. Analisis ini bertujuan untuk melihat secara keseluruhan tentang kualitas tes.

Dari hasil analisis diperoleh reliabilitas tes untuk paket soal 1 sebesar 0,756. Menurut Feldt dan Brehmman mengatakan bahwa suatu instrumen yang memiliki koefisien reliabilitas r ≥0,7 sudah dikatakan reliabel. tes untuk paket soal 1 bisa dipercaya penggunaannya.

Berdasarkan hasil analisis IRT (ITEMAN), maka dapat disimpulkan tingkat kesukaran (*Difficulty Item Index*)sebagai berikut :

Tabel 3

Kesimpulan Tingkat Kesukaran Butir Soal Paket Soal 2

Butir Soal	Keterangan
1	Mudah
2	Sedang
3	Sedang
4	Sedang
5	Mudah
6	Mudah
7	Sedang
8	Sedang
9	Sedang
10	Mudah
11	Mudah
12	Mudah
13	Sedang
14	Sukar
15	Sedang
16	Sedang
17	Sedang
18	Mudah
19	Mudah
20	Sedang

Tabel 3 di atas menujukkan bahwa tingkat kesukaran soal paket 2(55%) dalam kategori sedang.

Berikut ini adalah kesimpulan kriteria baik tidaknya butir soal paket soal 2 :

Tabel 4

Rekapitulasi Kriteria Baik Tidaknya Butir Soal Paket Soal 2

Butir Soal	Kualitas Butir
1	sangat baik
2	Baik
3	Dibuang
4	sangat baik
5	sangat baik
6	sangat baik
7	Dibuang
8	Baik
9	sangat baik
10	sangat baik
11	sangat baik
12	sangat baik
13	Dibuang
14	Baik
15	Baik

16	Baik
17	sangat baik
18	sangat baik
19	sangat baik
20	Dibuang

Butir soal pada paket soal 2 yang didrop atau tidak dipergunakan yaitu butir soal 3,7,13dan 20. Sehingga hanya 16 butir soal yang memenuhi standar atau baku.

Dari hasil analisis diperoleh reliabilitas tes untuk paket soal 2 sebesar 0,763. Artinya tes untuk Paket soal 2 reliabel atau konsisten, sehingga penggunaannya bisa dipercaya.

Berdasarkan hasil analisis IRT (ITEMAN), maka dapat disimpulkan tingkat kesukaran (*Difficulty Item Index*)sebagai berikut :

Tabel 5

Kesimpulan Tingkat Kesukaran Butir Soal Paket Soal 3

Butir Soal	**Keterangan**
1	Sukar
2	Sukar
3	Sedang
4	Sedang
5	Mudah
6	Mudah
7	Sedang
8	Sedang
9	Sukar
10	Sukar
11	Sukar
12	Sukar
13	Sedang
14	Sukar
15	Sedang
16	Sedang
17	Sedang
18	Mudah
19	Sukar
20	Sedang

Tabel 5 di atas menujukkan bahwa tingkat kesukaran soal paket 3 (45 %) dalam kategori sedang.

Berikut ini adalah kesimpulan kriteria baik tidaknya butir soal paket soal 3 :

Tabel 6

Rekapitulasi Kriteria Baik Tidaknya Butir Soal Paket Soal 2

Butir Soal	**Kualitas Butir**
1	Dibuang
2	Baik
3	Sangat Baik
4	Dibuang
5	Dibuang
6	sangat baik
7	sangat baik
8	sangat baik
9	Dibuang
10	sangat baik
11	Baik
12	sangat baik
13	sangat baik
14	sangat baik
15	sangat baik
16	Dibuang
17	sangat baik
18	Dibuang
19	Dibuang
20	Dibuang

Butir soal pada paket soal 3 yang didrop atau tidak dipergunakan yaitu butir soal 1,4,5,9,18,19 dan 20. Sehingga hanya 12 butir soal yang memenuhi standar atau baku.

Dari hasil analisis diperoleh reliabilitas tes untuk paket soal 3 sebesar 0,584. Artinya tes untuk paket soal 3 kurang reliabel atau kurang konsisten, sehingga penggunaannya kurang bisa dipercaya.

Setelah semua paket soal dianalisis kualitasnya, maka soal dipilah kembali disesuaikan dengan indikatornya dan kualitas soalnya disamakan atau relatif disamakan antar paket soal yang disediakan. Setalah dikalibrasi maka butir soal yang dipergunakan adalah 15 soal untuk penelitian ini.

Tes Mid Semester diberikan tiga paket soal dengan kualitas yang sama pada kelompok rombongan belajar mahasiswa mata kuliah *methodology of research*.

Deskripsi hasil mid semester pada mata kuliah methodology of research untuk tiga

kelompok rombongan belajar dengan paket soal yang berbeda, disajikan sebagai berikut :

Tabel 7

Deskripsi Hasil Tes Mid Semester tiga kelompok belajar mahasiswa pada mata kuliah methodology of research

Descriptives

PAKET SOAL

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum
					Lower Bound	Upper Bound		
1	19	50,53	23,799	5,460	39,06	62,00	20	90
2	26	36,73	10,857	2,129	32,35	41,12	15	70
3	10	49,50	15,890	5,025	38,13	60,87	25	75
Total	55	43,82	18,206	2,455	38,90	48,74	15	90

Mean pada kelompok rombongan belajar pertama (50,53) dengan standar deviasi 23, 799. Lebih besar dari kelompok rombongan belajar kedua (35,73) dan kelompok rombongan belajar ketiga (43,82). Penyebaran data pada kelompok belajar pertama relatif heterogen. Selain itu data di atas menggambarkan bahwa hasil tes mid semester mahasiswa pada ketiga kelompok rombongan belajar yang mengambil mata kuliah *methodology of research* relatif rendah, karena rata-ratanya di bawah 60.

Untuk memenuhi persyaratan uji perbedaan, maka data terlebih dahulu diuji kenormalannya. Berikut ini hasilnya :

Tabel 8

Uji Normalitas

Tests of Normality

		Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	HASIL TES	Statistic	df	Sig.	Statistic	df	Sig.
PAKET SOAL	1	,163	18	,200*	,903	18	,066
	2	,273	24	,000	,918	24	,053
	3	,135	10	,200*	,972	10	,904

*. This is a lower bound of the true significance.

a. Lilliefors Significance Correction

Berdasarkan uji normalitas ketiga kelompok dan mengacu pada hipotesis uji normalitas, yaitu :

H_o : Data berdistribusi normal

H_1 : Data berdistribusi tidak normal

Maka data ketiga kelompok tersebut untuk paket soal 1, paket soal 2 dan paket soal 3, yaitu : Sig pada kelompok paket soal 1 *p value* = 0,066 >*p value* = 0,05, Sig pada kelompok paket soal 2 *p value* = 0,053 >*p value* = 0,05, Sig pada kelompok paket soal 3 *p value* = 0,94 >*p value* = 0,05

Berarti Ho diterima (*rejected fail*). Artinya ketiga kelompok dengan tiga paket soal berasal dari populasi berdistribusi normal.

Uji homogenitas bertujuan untuk mengetahui apakah ketiga kelompok yang mengerjakan tes dengan paket soal 1, 2 dan paket soal 3 berasal dari populasi yang sama.

Di bawah ini disajikan tabel uji homogenitas:

Tabel 9

Uji Homogenitas

Test of Homogeneity of Variances

HASIL TES

Levene Statistic	df1	df2	Sig.
1,770	9	29	,118

Mengacu pada hipotesis uji homogenitas sebagaimana berikut H_o : Data berasal dari populasi yang sama, H_1 : Data bukan dari populasi yang sama

Berdasarkan uji homogenitas, diperoleh sig = 0,118 >*p value* = 0,05. Maka Ho diterima (*rejected fail*), artinya ketiga kelompok berasal dari populasi yang sama (homogen).

c. Uji Perbedaan Hasil Tes Berdasarkan Paket Soal

Untuk melihat apakah ada perbedaan hasil tes antara ketiga kelompok yang mengerjakan tes dengan paket soal yang berbeda dan dengan kualitas yang sama, maka dilakukan analisis uji statistik *one way anova*. Di bawah ini hasil analisis *one way anova* (output SPSS) :

Tabel 10

Uji Perbedaan

ANOVA

HASIL TES

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	6,832	14	,488	,771	,690
Within Groups	18,350	29	,633		
Total	25,182	43			

Dari tabel anova diperoleh sig = 0,690 > p value = 0,05. Maka Ho *rejected fail*. Artinya tidak ada perbedaan hasil tes diantara tiga kelompok dengan menggunakan paket soal yang berbeda tetapi kualitas soal sama.

Pengembangan tes berbasis komputer (CBT) dengan menggunakan software lebih praktis dan efisien. Bagi peserta tes khususnya mahasiswa STAIN Parepare yang tidak terbiasa dengan tes berbasis komputer awalnya mengalami sedikit kesulitan, terutama bagi peserta tes yang gagap teknologi. Karena itu setting pengembangan tes ini didesain semudah mungkin, salah satu desain yang relatif mudah untuk dipahami adalah *software easy quiz*.

Pada kalibrasi tes pada setiap paket soal dengan anaisis tingkat kesukaran, validitas dan reliabilitas secara umum menunjukkan butir-butir soal relatif bisa dipercaya penggunaannya. Hal ini karena penulis sebelum paket soal dikalibrasi atau diujicobakan secara empirik, penulis melakukan validisasi secara konsep dengan melakukan pendekatan kualitatif dengan menanyakan kepada teman sejawat yang mengampu mata kuliah *methodology of research*.

Tes dengan komputer yang sudah dikalibrasi, selanjutnya dimanfaatkan pada skala terbatas. Dalam hal ini digunakan pada mid tes mata kuliah *methodology of research*. Tiga paket soal yang sudah dikalibrasi diteskan kepada tiga kelompok belajar yang sedang mengikuti mata kuliah *methodology of research*. Pada tes ini tidak semua testee diambil untuk dianalisis, hanya sebagian yang dianalisis yaitu sampel yang terkena random. Ini dilakukan untuk memenuhi salah satu persyaratan uji inferensial dan untuk mengeneralisasi hasil. Pada hasil mid tes, ternyata tidak ada perbedaan hasil tes dengan paket soal 1, paket soal 2 dan paket soal 3 dengan kualitas soal yang sama. Ini menunjukkan bahwa kemampuan tiga kelompok tes mempunyai kemampuan yang sama atau paket soal yang diberikan tidak mempengaruhi perbedaan kemampuan testee. Sehingga bisa dikatakan kualitas soal benar-benar relatif sama.

SIMPULAN

Berdasarkan hasil penelitian dan pembahasan yang dikemukakan oleh penulis, maka pada penelitian ini dapat disimpulkan sebagai berikut : Pengembangan tes dapat dilakukan dengan menggunakan komputer (*computer base test*). Penggunaan komputer sebagai pengganti tes yang menggunakan *paper and pencil* lebih efisien dan efektif, Perakitan tes dengan komputer dapat menggunakan software yang sudah ada, salah satunya *software easy quiz*. Pada praktiknya penggunaan software easy quiz dalam tes relatif tidak terlalu sulit untuk dipahami, karena settingnya relatif sederhana. Standarisasi tes pada mid tes mata kuliah *methodology of research* dilakukan dengan melakukan analisis yaitu tingkat kesukaran, validitas dan reliabilitas. Hal ini perlu dilakukan agar kualitas tes terjamin, yang kemudian dijadikan sebagai masukkan untuk bank soal. Hasil analisis butir tes menujukkan bahwa untuk paket soal 1 tingkat kesukaran paket soal 1(85%) dalam kategori sedang, 15 butir soal yang digunakan, reliabilitas tes untuk paket soal 1 sebesar 0,756 artinya tes untuk paket soal 1 bisa dipercaya penggunaannya. Paket soal 2, tingkat kesukaran soal paket 2(55%) dalam kategori sedang, 16 butir soal yang digunakan, reliabilitas tes untuk paket soal 2 sebesar 0,763. Artinya tes untuk Paket soal 2 reliabel atau konsisten, sehingga penggunaannya bisa dipercaya. Untuk paket soal 3 tingkat kesukaran soal paket 3 (45 %) dalam kategori sedang, reliabilitas tes untuk paket soal 3 sebesar 0,584. Artinya tes untuk paket soal 3 kurang reliabel atau kurang konsisten, sehingga penggunaannya kurang bisa dipercaya. Namun setelah diperbaiki butir soal yang dibuang maka reliabilitasnya pun jadi meningkat. Pada uji perbedaan mid tes untuk ketiga kelompok belajar mata kuliah methodology of research dengan paket soal yang sudah dikalibrasi diperoleh uji one way anova sebesar F hitung = 0,771, dan signifikansi 0,690.

Karena sig > p value = 0,05. Artinya tidak ada perbedaan hasil mid tes pada ketiga kelompok belajar mata kuliah *methodology of research*.

Daftar Pustaka

Anderson, O.W dan Karthwohl. D. R, 2001. *A Taxonomy For Learning Teaching and Assessing (A Revision of Blooms Taxonomy of Educational Objektives),* New York: Addition Wesley, Longman.

Agustina. 2010. *Tutorial 5 Hari Menguasai Adobe Flash CS4*. Semarang: Andi

Azwar Syarifuddin, 2003. *Tes Prestasi: Fungsi dan Pengembangan Pengukuran Prestasi Belajar*, edisi II, cetakan ke 4: Pustaka Pelajar.

Bitzer, D. L. (2000). *A comparative analysis of web based testing and evaluation systems*. diakses 10 Mei 2013,http://renoir.csc.nscu.edu/MRA/Reports/WebBasedTesting.html.

Cassady, J. C. & Gridley, B. E, (2005). The effects of online formative and summative assessment on test anxiety and performance. *Journal of Technology, Learning, and Assessment, 4*(1). Diambil 12 Mei 2013 dari http://www.jtla.org.

Crocker, L & Algina, J (1986). *Introduction to Classical and Modern Test Theory*. New York : Holt, Rinehart and Winston, Inc.

Hambleton, Ronald K; Swaminathan, H ; and Rogers, H. Jane (1991). *Fundamentals of Item Response Theory California* : Sage Publications, The International Profesional Publishers.

Hidayat, Wahyu. (2012). *Evaluasi Pembelajaran PAI*. Yogyakarta: Gre Publishing

Linn, R.L, Grondlund, N.E. (2000). *Measurement and Assessment In Teaching* . Eighth edition. New Jersey: Merril an imprint of Prentice Hall.

Russell, M. (1999). Testing on computers: A follow-up study comparing performance on computer and on paper. *Education Policy Analysis Archives*, *7*(20). Diakses dari http://epaa.asu.edu/epaa/v7n20.

Safari. (2005). *Teknik Analisis Butir Soal Instrumen Tes dan Non-Tes*. Jakarta : Asosiasi Pengawas Sekolah Indonesia Departemen Pendidikan Nasional

Sapriati, Amalia dan Minrohayati ujian berbasis komputer (ubk).*Jurnal Pendidikan Terbuka dan Jarak Jauh, Volume 10, Nomor 2, September 2009, 63-72*.

Shepherd, E. (2003). Delivering computerized assessments safely and securely. *The eLearning Developers' Journal*. Diambil 20 Oktober 2003, dari http://www.eLearningGuild.com.

Sudjana, N. (1989). *Penilaian Hasil Belajar Mengajar*. Bandung : PT. Remaja Rosdakarya.

Sudijono Anas, 2005. *Pengantar Evaluasi Pendidikan*, Jakarta: Rajawali Press.

Test, measurement & Research Service. www.PearsonAssessments.com, diunduh pada 11 Mei 2013.