PENGEMBANGAN DAN ANALISIS SOAL ULANGAN KENAIKAN KELAS KIMIA SMA KELAS X BERDASARKAN *CLASSICAL TEST THEORY* DAN *ITEM RESPONSE THEORY*

Nahadi, Wiwi Siswaningsih, dan Ana Rofiati

Jurusan Pendidikan Kimia, FPMIPA
Universitas Pendidikan Indonesia

ABSTRACT

This research is title "Test Development and Analysis of First Grade Senior High School Final Examination in chemistry Based on Classical Test Theory and Item Response Theory". This research is conducted to develop a standard test instrument for final examination in senior high school at first grade using analysis based on classical test theory and item response theory. The test is a multiple choice test which consists of 75 items. Each item has five options. The research method is research and development method to get a product of test items which fulfill item criterion such as validity, reliability, item discrimination, item difficulty and distracting options quality based on classical test theory and validity, reliability, item discrimination, item difficulty and pseudo-guessing based on item response theory. The three parameter item response theory model is used in this research. Research and development method is conducted until preliminary field test to 102 first grade students in senior high school. Based on the research result, the test fulfills criterion as a good instrument based on classical test theory and item response theory. The final examination test items have vary of item quality so that some of them need a revision to make them better either for the stem and the options. From the total of 75 test items, 21 test items are declined and 54 test items are accepted.

PENDAHULUAN

Pembelajaran adalah proses interaksi peserta didik dengan guru dan sumber belajar pada suatu lingkungan belajar. Proses pembelajaran perlu direncanakan, dilaksana-kan, dinilai, dan diawasi agar terlaksana secara efektif dan efisien. Dalam sistem pembelajaran, evaluasi merupakan salah satu komponen penting dan tahap yang harus ditempuh oleh guru untuk mengetahui keefektifan pembelajaran (Arifin, 2009). Menurut Peraturan Pemerintah Nomor 19 Tahun 2005 Tentang Standar Nasional Pendidikan, penilaian adalah proses pengumpulan dan pengolahan informasi untuk mengukur pencapaian hasil belajar peserta didik. Penilaian hasil belajar oleh pendidik dilakukan secara berkesinambungan untuk memantau proses, kemajuan, dan perbaikan hasil dalam bentuk ulangan harian, ulangan tengah semester, ulangan akhir semester dan ulangan kenaikan kelas. Penilaian digunakan untuk menilai pencapaian kompetensi peserta didik, bahan penyusunan laporan kemajuan hasil belajar dan memperbaiki proses pembelajaran.

Ulangan kenaikan kelas sebagai bentuk evaluasi terutama untuk tingkat SMA kelas X memiliki fungsi penting dalam proses seleksi peserta didik untuk penentuan program penjurusan yang terdiri dari Ilmu Alam, Ilmu Sosial atau Bahasa. Dalam proses evaluasi pembelajaran atau penilaian proses dan hasil belajar, guru sering menggunakan instrumen tertentu. Untuk mengetahui suatu instrumen atau tes yang digunakan termasuk baik atau kurang baik, maka perlu dilakukan analisis butir soal (*Item analysis*). Dengan analisis butir soal dapat diperoleh informasi tentang kejelekan sebuah soal dan petunjuk untuk mengadakan perbaikan (Arikunto, 2010). Dalam teori pengukuran, terdapat dua model pengukuran, yaitu *classical test theory* dan *item response theory* (Courville, 2004).

Menurut Hambleton (Gleason, 2008) *classical test theory* telah digunakan selama bertahun-tahun untuk menentukan tingkat kesukaran dan karakteristik lainnya dalam instrumen pengukuran. Namun, ada beberapa kekurangan dalam *classical test theory*. Kekurangan yang paling menonjol dalam teori ini adalah indeks butir soal seperti tingkat kesukaran dan daya pembeda yang didapatkan

dengan menggunakan *classical test theory* bergantung pada kelompok peserta tes dan penilaian kemampuan peserta tes bergantung pada pemilihan butir soal instrumen. Para ahli psikometri telah mengembangkan teori pengukuran baru yang disebut *item response theory* (IRT) dalam rangka mengatasi kekurangan tersebut.

Hambleton *et al.* (1991) mengemukakan *item response theory* (IRT) adalah teori yang menyatakan bahwa hasil tes dapat diprediksikan atau dijelaskan melalui serangkaian faktor yang disebut dengan sifat atau karakter (*trait*), karakter terpendam (*latent trait*) atau kemampuan (*abilities*) dan hubungan antara jawaban peserta tes dengan kemampuannya dapat dijelaskan dengan grafik karakteristik butir atau *item characteristic curve* (ICC). Semakin tinggi tingkat kemampuan, semakin besar peluang jawaban benar dari suatu butir soal. Sehingga dapat dikatakan bahwa *classical test theory* adalah "*test based*" atau berorientasi pada keseluruhan tes sedangkan IRT merupakan "*item based*" atau lebih berorientasi pada tiap butir soal.

Berdasarkan pemaparan di atas, maka dilakukan penelitian untuk mengembangkan butir soal ulangan kenaikan kelas dengan menerapkan *classical test theory* dan *item response theory* dalam tahap analisis soal. Permasalahan utama pada penelitian ini adalah "Apakah bentuk soal ulangan kenaikan kelas kimia SMA Kelas X yang dikembangkan telah memenuhi kriteria sebagai instrumen yang baik berdasarkan *classical test theory* dan *item response theory*?". Tujuan dari penelitian ini adalah untuk mengembangkan dan menganalisis soal ulangan kenaikan kelas kimia SMA kelas X pada semester genap kesukaran berdasarkan *classical test theory* dan *item response theory*.

Classical Test Theory

Classical Test Theory (CTT-Teori Tes Klasik), yang lebih dikenal sebagai CTT, dikembangkan sekitar tahun 1920-an (Natarajan, 2009). Teori ini memiliki beberapa komponen seperti teori validitas, reliabilitas, objektivitas, teori analisis tes, teori analisis butir dan sebagainya. Sebagian besar praktiknya dimulai dari tes psikologi dan kemudian dikembangkan dalam tes kependidikan. Teori tes klasik menurut O'Connor *et al.* (2002) adalah suatu model pengukuran berdasarkan informasi yang didapatkan pada level skor tes. Menurut Hambleton dan Jones, teori tes klasik adalah teori mengenai skor tes yang mengenalkan tiga konsep yaitu *test score/observed score*, *true score* dan *error score*. Berikut ini parameter yang digunakan dalam teori tes klasik.

1. Tingkat Kesukaran

Perhitungan tingkat kesukaran soal adalah pengukuran seberapa besar derajat kesukaran suatu soal. Jika suatu soal memiliki tingkat kesukaran seimbang (proporsional), maka dapat dikatakan bahwa soal tersebut baik. Suatu soal tes hendaknya tidak terlalu sukar dan tidak pula terlalu mudah (Arifin, 2009). Menurut Arikunto (2010), soal yang terlalu mudah tidak merangsang peserta didik untuk mempertinggi usaha memecahkannya. Sebaliknya soal yang terlalu sukar akan menyebabkan peserta didik menjadi putus asa dan tidak mempunyai semangat untuk mencoba lagi karena di luar jangkauannya.

2. Daya Pembeda

Perhitungan daya pembeda adalah pengukuran sejauh mana suatu butir soal mampu membedakan peserta didik yang sudah menguasai kompetensi dengan peserta didik yang belum/kurang menguasai kompetensi berdasarkan kriteria tertentu. Semakin tinggi koefisien daya pembeda suatu butir soal, semakin mampu butir soal tersebut membedakan antara peserta didik yang menguasai kompetensi dengan pesserta didik yang kurang menguasai kompetensi (Arifin, 2009).

3. Kualitas Pengecoh

Pada soal bentuk pilihan ganda ada alternatif jawaban (opsi) yang merupakan pengecoh. Butir soal yang baik, pengecohnya akan dipilih secara merata oleh peserta didik yang menjawab salah. Sebaliknya butir soal yang kurang baik, pengecohnya akan dipilih secara tidak merata (Arifin, 2009).

Menurut Firman (2000) analisis pengecoh bertujuan untuk menemukan

pengecoh yang kurang berfungsi dengan baik. Arikunto (2010) mengemukakan bahwa suatu pengecoh dapat dikatakan berfungsi baik jika paling sedikit dipilih oleh 5% pengikut tes.

Item Response Theory

Hambleton *et al.* (1991) mengemukakan bahwa *item response theory* (IRT) adalah teori yang menyatakan bahwa hasil tes dapat diprediksikan atau dijelaskan melalui serangkaian faktor yang disebut dengan sifat atau karakter (*trait*), karakter terpendam (*latent trait*) atau kemampuan (*abilities*) dan hubungan antara jawaban peserta tes dengan kemampuannya dapat dijelaskan dengan grafik karakteristik butir atau *item characteristic curve* (ICC).

Dengan menggunakan IRT, kemampuan peserta tes dapat dievaluasi dan seberapa baik kemampuan suatu butir soal dalam suatu tes dapat dideskripsikan (Act Workforce Development, 2010). IRT menggunakan konsep Item Characteristic Curve (ICC-Kurva Karakteristik Butir) untuk menunjukkan hubungan antara kemampuan peserta tes dengan kemampuan butir soal. Dalam IRT, kemampuan peserta tes dengan parameter butir dapat diamati berdasarkan pola respon peserta tes pada suatu tes. Banyaknya parameter butir yang diamati menentukan model IRT yang mana yang akan digunakan. Meskipun model-model ini melibatkan prosedur matematis yang rumit, tetapi konsep dasarnya mudah dipahami.

Parameter butir merupakan konsep dasar dari IRT. Secara umum model-model IRT didasarkan pada satu, dua dan tiga parameter. Berikut ini adalah model IRT yang didasarkan pada tiga parameter.

1. Parameter *a*: Daya Pembeda

Salah satu ciri tes yang baik adalah peserta tes pada kelompok atas akan memiliki pilihan jawaban benar lebih banyak daripada peserta tes pada kelompok bawah. Parameter *a* menunjukkan seberapa baik sebuah butir soal dapat membedakan peserta tes dengan tingkat kemampuan yang berbeda-beda. Butir soal yang baik biasanya memiliki rentang nilai daya pembeda dari 0,5 sampai 2. Hal ini digambarkan dengan plot grafik ICC. Semakin tinggi kemiringan suatu ICC, semakin tinggi pula daya pembeda suatu butir soal. Daya pembeda yang tinggi menunjukkan bahwa peserta tes yang memiliki skor tinggi cenderung menjawab butir soal dengan benar, sedangkan peserta tes dengan skor rendah cenderung memilih pilihan jawaban yang salah.

2. Parameter *b*: Tingkat Kesukaran

Tingkat kesukaran sebuah butir soal, dikenal sebagai parameter *b*, adalah titik dimana kurva bentuk-S memiliki kemiringan paling tinggi. Semakin sukar suatu butir soal, semakin tinggi kemampuan yang diperlukan dari peserta tes untuk menjawab butir soal tersebut dengan benar. Suatu butir soal dengan nilai *b* yang tinggi adalah soal sukar, dimana peserta tes kelompok bawah tidak dapat menjawab dengan benar. Butir soal dengan nilai *b* rendah adalah soal mudah, dimana sebagian besar peserta tes, termasuk peserta tes kelompok bawah, akan memiliki peluang minimal setengah untuk menjawab soal tersebut dengan benar.

3. Parameter *c*: Faktor Tebakan

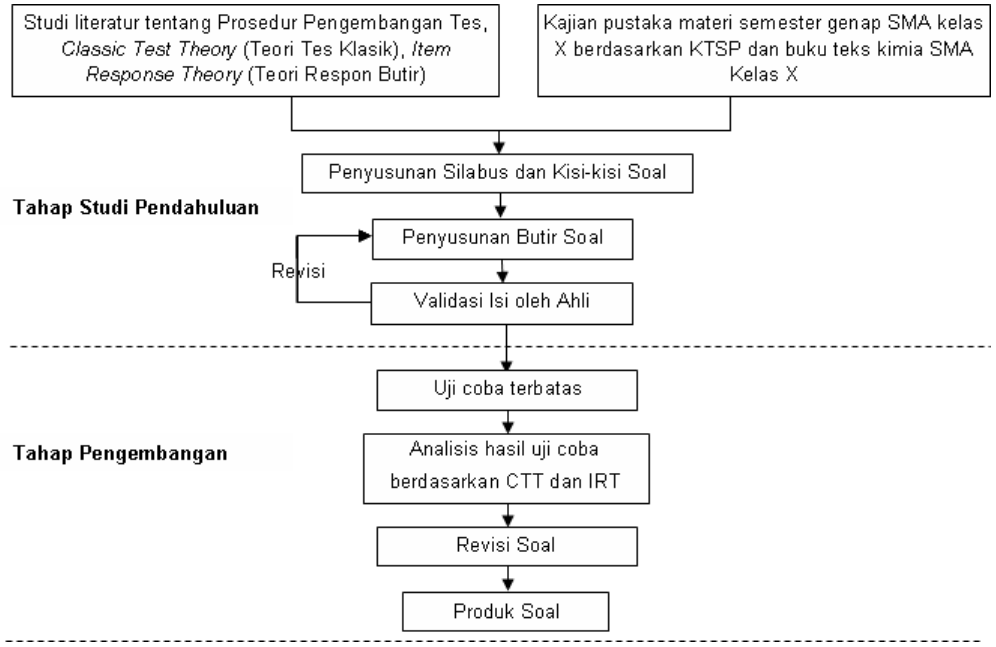
Beberapa model IRT memasukkan parameter faktor tebakan. Parameter *c* menunjukkan kecenderungan peserta tes kelompok bawah dapat menebak jawaban yang benar dari sebuah butir soal sehingga memiliki peluang lebih dari nol dalam menjawab soal dengan benar. Sebagai contoh, peserta tes yang memilih jawaban secara acak dari sebuah butir soal yang memiliki empat pilihan respon jawaban dapat menjawab butir soal dengan benar satu dari empat kali kesempatan, artinya peluang menjawab benar dengan menebak adalah 0,25. ICC masih berbentuk-S, tetapi nilai terendah dari kurva adalah lebih dari nol.

METODE

Dalam penelitian ini metode penelitian yang digunakan adalah metode penelitian dan pengembangan. Sukmadinata (2010) mengemukakan bahwa terdapat tiga tahap dalam pelaksanaan penelitian dan pengembangan, yaitu (1) Studi pendahuluan yang meliputi studi literatur, studi lapangan dan penyusunan draf awal produk, (2)

Pengembangan model yang meliputi uji coba dengan sampel terbatas (uji coba terbatas) dan uji coba dengan sampel lebih luas (uji coba lebih luas), (3) Uji model yang meliputi uji produk melalui eksperimen dan sosialisasi produk.

Penelitian ini dilakukan sampai pada tahap uji coba terbatas pada langkah pengembangan model. Tujuan dari penelitian dan pengembangan ini adalah untuk menghasilkan produk evaluasi berupa soal ulangan kenaikan kelas. Hal yang akan dideskripsikan adalah mengenai kualitas soal yang dikembangkan berdasarkan *Classical Test Theory* (CTT-Teori Tes Klasik) dan *Item Response Theory* (IRT-Teori Respon Butir).



Pada dasarnya penelitian ini dilakukan melalui dua tahap yaitu tahap studi pendahuluan dan tahap pengembangan. Tahapan pertama dalam penelitian ini adalah studi pendahuluan yang merupakan tahap persiapan untuk pengembangan. Tahap ini dilakukan dengan studi kepustakaan dan penyusunan produk awal. Tahap kedua dalam penelitian ini adalah tahap pengembangan yaitu uji coba produk. dalam penelitian ini, uji coba pengembangan produk dilakukan sampai pada tahap uji coba terbatas. Sampel uji coba instrumen adalah 102 siswa Kelas X SMA Negeri di kota Bandung. Hasil uji coba terbatas dianalisis berdasarkan *classical test theory* dan *3PL model-item response theory*, kemudian hasil analisisnya digunakan sebagai bahan pertimbangan untuk revisi produk soal.

HASIL DAN PEMBAHASAN

Secara keseluruhan soal ulangan kenaikan kelas yang dikembangkan memenuhi kriteria sebagai tes yang baik. Hal ini didasarkan pada hasil uji validitas isi, validitas soal dan reliabilitas tes. Hasil uji validitas isi menyatakan semua soal sebanyak 75 butir di dalam tes adalah valid. Hasil validasi isi ini menunjukkan bahwa seluruh soal telah valid, yang artinya soal telah sesuai dengan standar kompetensi, kompetensi dasar dan indikator pembelajaran yang akan diukur serta telah sesuai dengan cakupan materi yang diujikan. Soal ulangan kenaikan kelas yang dikembangkan kemudian diujicobakan untuk mengetahui kualitas tiap butir soal maupun kualitas tes secara keseluruhan. Setelah diuji validitas dan reliabilitas tes, dapat dinyatakan bahwa soal ulangan kenaikan kelas yang dikembangkan memenuhi syarat

tes yang baik karena memiliki nilai validitas dan reliabilitas yang sangat tinggi. Perhitungan validitas berdasarkan program anates menghasilkan nilai validitas soal sebesar 0,83, yang menunjukkan bahwa soal ini memiliki validitas yang sangat tinggi. Berdasarkan program anates maupun xcalibre dihasilkan nilai reliabilitas tes sebesar 0,91 yang menunjukkan bahwa soal ini memiliki reliabilitas sangat tinggi.

Kualitas butir soal berdasarkan *classical test theory* dan *item response theory*, ditinjau dari validitas, daya pembeda, tingkat kesukaran dan kualitas pengecoh serta faktor tebakan tiap butir soal, memiliki nilai yang bervariasi sehingga dari 75 butir soal, ada butir-butir soal yang memenuhi kriteria butir soal yang baik sehingga dapat diterima, ada butir soal yang dapat diterima namun memerlukan revisi dan beberapa butir soal ditolak atau dibuang.

Analisis kualitas tiap butir soal dilakukan berdasarkan 2 teori yaitu, *classical test theory* (CTT) dan *item response theory* (IRT). Berdasarkan *classical test theory* dilihat dari segi validitas butir soal, daya pembeda dan tingkat kesukaran serta kualitas pengecoh. Sedangkan analisis kualitas butir soal berdasarkan *item response theory* ditinjau dari segi validitas-S, validitas-T, reliabilitas, daya pembeda (parameter *a*), tingkat kesukaran (parameter *b*) dan faktor tebakan (parameter *c*).

Butir soal nomor 1, 35, 37, 41 dan 50 memiliki validitas sangat rendah. Berdasarkan analisis CTT, dapat diketahui bahwa daya pembeda butir ini jelek dan termasuk butir soal yang mudah serta memiliki kualitas pengecoh yang kurang baik. Berdasarkan analisis IRT, butir-butir ini tidak terkalibrasi atau tidak dapat dianalisis, hal ini dikarenakan hasil analisis untuk butir-butir ini berada di luar kriteria butir tes yang dapat diterima sehingga berdasarkan analisis IRT tidak dapat diketahui daya pembeda, tingkat kesukaran maupun faktor tebakannya. Kedua hasil analisis tersebut menunjukkan hasil yang tidak jauh berbeda. Ditinjau dari hasil analisis kedua teori tersebut, butir soal nomor 1, 35, 37, 41 dan 50 ditolak atau dibuang.

Butir soal nomor 7 memiliki validitas yang tinggi baik menurut analisis CTT maupun IRT. Butir ini juga memiliki reliabilitas yang sangat tinggi. Berdasarkan analisis CTT (gambar 1,2,3), butir ini memiliki daya pembeda yang baik sekali dengan kategori tingkat kesukaran soal yang sedang. Seluruh pengecoh pada butir ini berfungsi dengan baik. Berdasarkan analisis IRT (gambar 4,5,6,7,8,9,10,11), butir ini memiliki daya pembeda yang sedang dengan kategori soal mudah dan faktor tebakan dapat diterima. Perbedaan interpretasi daya pembeda dan tingkat kesukaran pada hasil analisis kedua teori tersebut disebabkan karena perbedaan parameter yaitu daya pembeda pada CTT menunjukkan selisih antara proporsi siswa kelompok atas yang menjawab benar dengan proporsi siswa kelompok bawah yang menjawab benar. Sedangkan pada IRT, daya pembeda menunjukkan kemiringan grafik IRF butir soal. Semakin curam kemiringan grafik IRF, semakin baik daya pembeda butir soal. Tingkat kesukaran menurut CTT merupakan hasil perhitungan proporsi siswa yang menjawab benar terhadap jumlah seluruh siswa. Sedangkan menurut IRT, tingkat kesukaran merupakan besarnya nilai dalam skala kemampuan atau θ yang dibutuhkan siswa untuk memiliki peluang menjawab soal dengan benar sebesar (1+c)/2. Pada butir soal nomor 7, untuk dapat memiliki peluang menjawab benar sebesar 0,65, siswa harus memiliki kemampuan sebesar 0,48. Berdasarkan hasil analisis kedua teori tersebut, butir soal nomor 7 dapat diterima.

Butir soal nomor 68 dan 69 merupakan butir soal yang baik menurut CTT maupun IRT karena kedua butir ini memiliki validitas yang tinggi dan daya pembeda yang baik. Kualitas seluruh pengecoh pada kedua butir soal ini berfungsi baik. Oleh karena itu, kedua butir ini diterima.

Sebanyak 16 butir soal lainnya yaitu butir soal nomor 3, 9, 10, 21, 22, 30, 31, 34, 42, 45, 53, 54, 56, 58, 62 dan 63 ditolak karena tidak memenuhi kriteria sebagai butir soal yang baik menurut *classical test theory*

maupun *item response theory*. Butir-butir ini memiliki validitas sangat rendah, sehingga dapat dikatakan bahwa skor butir-butir ini sangat kurang memiliki korelasi dengan skor totalnya dan dengan *theta* atau kemampuan siswa.

Sebanyak 51 butir lainnya memiliki kualitas butir yang bervariasi dengan sebagian besar butir memiliki validitas dengan kategori cukup, tinggi dan sangat tinggi baik menurut *classical test theory* maupun *item response theory*. Beberapa butir tetap memerlukan revisi pada pengecohnya karena dari empat pilihan jawaban yang disajikan, tidak semua dipilih oleh 5% pengikut tes sehingga perlu diadakan perbaikan kualitas pengecoh. Sedangkan butir-butir soal yang memiliki validitas cukup baik namun daya pembedanya kurang, revisi dilakukan pada pokok soal untuk memperbaiki kualitas daya pembedanya.

Berdasarkan pembahasan di atas, dapat dikatakan bahwa hasil analisis CTT dan IRT secara umum untuk tiap butir soal tidak jauh berbeda. Namun terdapat pula beberapa hasil analisis pada butir soal yang memiliki perbedaan signifikan.

KESIMPULAN DAN SARAN

1. Kesimpulan

Berdasarkan hasil penelitian dan pembahasan, soal ulangan kenaikan kelas yang dikembangkan memenuhi kriteria sebagai alat ukur yang baik, sehingga dapat disimpulkan bahwa:

a. Soal ulangan kenaikan kelas yang dikembangkan memiliki validitas isi dan validitas empiris yang memenuhi kriteria sebagai tes yang baik berdasarkan *classical test theory* dan *item response theory*.

b. Soal ulangan kenaikan kelas yang dikembangkan telah memenuhi kriteria sebagai tes yang baik dilihat dari reliabilitasnya.

c. Soal ulangan kenaikan kelas yang dikembangkan telah memenuhi kriteria sebagai tes yang baik dilihat dari reliabilitasnya.

d. Soal ulangan kenaikan kelas yang dikembangkan memiliki tingkat kesukaran yang memenuhi kriteria sebagai tes yang baik berdasarkan *classical test theory* dan *item response theory*.

e. Soal ulangan kenaikan kelas yang dikembangkan memiliki daya pembeda yang memenuhi kriteria sebagai tes yang baik berdasarkan *classical test theory* dan *item response theory*.

f. Soal ulangan kenaikan kelas yang dikembangkan memiliki kualitas pengecoh yang memenuhi kriteria sebagai tes yang baik berdasarkan *classical test theory*.

g. Soal ulangan kenaikan kelas yang dikembangkan memiliki faktor tebakan yang memenuhi kriteria sebagai tes yang baik berdasarkan *item response theory*.

2. Saran

Setelah melakukan penelitian ini, peneliti menyarankan agar:

a. Pengembangan tes dengan memperhatikan aturan-aturan pembuatan tes yang baik seperti uji coba dan analisis kualitas tes sebaiknya lebih sering diterapkan di sekolah-sekolah pada saat proses penyusunan alat evaluasi sehingga soal yang diberikan benar-benar dapat mengukur kemampuan siswa.

b. Tes yang dikembangkan sebaiknya memiliki penyebaran tingkat kesukaran yang seimbang (proporsional) untuk kategori soal mudah, sedang dan sukar.

c. Lebih banyak lagi peneliti yang mengembangkan tes berdasarkan *item response theory* sehingga teori ini dapat lebih dikenal dan dapat digunakan aplikasinya secara lebih luas.

DAFTAR PUSTAKA

Act Workforce Development. (2010). *Introduction to Test Development for Credentialing: Item Response Theory*. Iowa city: Act, Inc.

Arifin, Z. (2009). *Evaluasi Pembelajaran*. Bandung: PT Remaja Rosdakarya.

Arikunto, S. (2010). *Dasar-dasar Evaluasi Pendidikan*. Jakarta: Bumi Aksara.

Courville, T.G. (2004). An Empirical Comparison of Item Response Theory and Classical Test Theory Item/Person Statistics. Disertasi pada Texas A&M University.

Firman, H. (2000). *Penilaian Hasil Belajar Dalam Pengajaran Kimia.* Bandung: Jurusan Pendidikan Kimia FPMIPA UPI.

Hambleton R.K, Swaminathan, H dan Rogers, H.J. (1991). *Fundamentals of Item Response Theory*. California: Sage Publications,Inc.

Natarajan. (2009). *Basic Principles of IRT and Application to Practical Testing & Assessment*. India: MeritTrac Services (P) Ltd.

O'Connor, L.G, Radcliff, C.J dan Gedeon, J.A. (2002). *Applying Systems Design and Item Response Theory to the Problem of Measuring Information Literacy Skills*. Kent University: College & Research Libraries.

Sukmadinata, N.S. (2010). *Metode Penelitian Pendidikan*. Bandung: PT Remaja Rosdakarya.