KARAKTERISTIK METODE PENYETARAAN SKOR TES UNTUK DATA DIKOTOMOS

Nonoh Siti Aminah
PMIPA FKIP UNS
nonoh_nst@yahoo.com

Abstrak

Penelitian ini bertujuan untuk menemukan: 1) akurasi estimasi parameter item pada *test equating* menggunakan metode *Item Characteristic Curve* (ICC). 2) sensitivitas metode *linear* yang terdiri atas *Tucker - Levine score method* dan *Levine true score method applied to observed scores* serta metode *equipercentile* yang terdiri atas metode *Braun-Holland linear* dan *chained equipercentile.* Data empiris yang digunakan yaitu respons siswa peserta Ulangan Akhir Semester V Mata Pelajaran Ilmu Pengetahuan Alam (IPA) SMP Tahun Ajaran 2009/2010. Penyetaraan tes menggunakan *anchor test design. Anchor test* bersifat *external, anchor test* berfungsi sebagai pengait antara tes yang disetarakan. Item *anchor* berisi 10 item materi Fisika. Banyak item pada tes A 55 item, tes B 55 item dan tes C 50 item. Pola penyetaraan yang digunakan pola kelompok, sehingga banyak item hasil penyetaraan berjumlah 140 item terdiri atas 10 *anchor item* milik bersama, 45 item berasal dari tes A, 45 item berasal dari tes B, dan 40 item berasal dari tes C. Hasil penelitian menunjukkan bahwa: 1) Estimasi parameter item pada penyetaraan tes menggunakan metode *Item Characteristic Curva* (ICC) menghasilkan formula indeks kesulitan item, 2) urutan sensitivitas metode penyetaraan dari paling tinggi sampai paling rendah yaitu *Tucker – Levine method, Levine method*, *Braun - Holland linear method. Chained Equipercentile Equating method.*

Kata kunci: *Test equating, anchor test, external anchor test*, *RMSD, RMSE*

THE CHARACTERISTICS OF TEST EQUATING METHODS FOR DICHOTOMOUS DATA

Nonoh Siti Aminah
PMIPA FKIP UNS
nonoh_nst@yahoo.com

Abstract

This study aims: 1) to find out the accuracy of item parameter estimates in test equating by means of the Item Characteristic Curve (ICC) method, and 2) to find out the sensivity of the linear methods consisting of the Tucker-Levine score method and the Levine true score method applied to observed scores and the equipercentile methods consisting of the Braun-Holland linear method the chained equipercentile equating method. The data were empirical data obtained from the response patterns of the junior high school students taking the final test of Natural Sciences in the odd semester of the academic year of 2009/2010. The test equating employed the external anchor test design. The anchor test served to unite the equated tests. The anchor test consisted of 10 physics items. The test A had 55 items, the test B had 55 items, and the test C had 50 items. The equating pattern employed the group pattern, so that in the equating there were 140 items, consisting of 10 common anchor items, 45 items from tests A, 45 items from tests B, and 40 items from tests C. The results of the study are as follows. 1) The item parameter estimate in the test score equating by means of the Item Characteristic Curve (ICC) method yields the formula for the item difficulty index, and 2) urutan sensitivitas metode penyetaraan dari paling tinggi sampai paling rendah yaitu The order of the sensitivity of the equating methods from the highest to the lowest is Tucker-Levine method, Levine method, Braun-Holland linear method. Chained Equipercentile Equating method.

Keywords: *test equating, anchor test, external anchor test, RMSD, RMSE*

Pendahuluan

Hasil pembelajaran di kelas berkaitan dengan pengetahuan, keterampilan dan sikap siswa. Pembelajaran yang berkualitas akan meningkatkan kualitas lulusan. Kualitas lulusan menjadi indikator dari kualitas pendidikan. Penilaian merupakan komponen penting dalam sistem pendidikan, sebab hasil penilaian mencerminkan perkembangan atau kemajuan hasil pendidikan yang dapat dibandingkan dari waktu ke waktu, antara satu sekolah dengan sekolah lainnya atau antara wilayah dengan wilayah lainnya. Proses penyamaan tingkat pencapaian hasil pendidikan antara sekolah atau antara wilayah dalam teori pengukuran disebut *equating* atau penyetaraan.

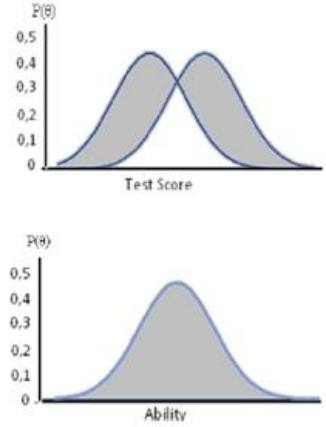
Ditinjau dari cakupannya, penilaian terdiri atas penilaian yang bersifat makro dan penilaian yang bersifat mikro. Penilaian yang bersifat makro cenderung menggunakan sampel untuk menelaah suatu program dan dampaknya, program yang dimaksud yaitu program pendidikan. Program pendidikan adalah program yang direncanakan untuk memperbaiki kualitas dalam bidang pendidikan. Penilaian yang bersifat mikro digunakan di tingkat kelas, yang bertujuan untuk mengetahui capaian hasil belajar, khususnya untuk mengetahui capaian hasil belajar siswa. Sasarannya yaitu program pembelajaran di kelas dengan penanggungjawab pengajar.

Capaian hasil belajar tidak hanya bersifat kognitif, melainkan juga mencakup semua potensi yang ada pada siswa. Penilaian pembelajaran diklasifikasikan menjadi dua, yaitu penilaian formatif dan penilaian sumatif. Penilaian formatif bertujuan untuk memperbaiki proses belajar mengajar. Hasil penilaian formatif dianalisis untuk mengetahui konsep yang belum dipahami sebagian siswa, kemudian diikuti kegiatan remedial. Kegiatan remedial yaitu kegiatan pembelajaran untuk mengatasi kesulitan belajar siswa yang diidentifikasi dari kegiatan penilaian formatif. Penilaian sumatif bertujuan menetapkan tingkat keberhasilan siswa.

Teknik penilaian antara lain observasi, penugasan, inventori, jurnal, penilaian diri dan penilaian yang sesuai dengan karakteristik kompetensi dan perkembangan siswa. Teknik penilaian dapat dilakukan dengan menggunakan teknik tes dan non tes. Tes merupakan alat pengumpul data

untuk mengetahui kemampuan individu atau kelompok dalam menyelesaikan suatu persoalan atau memperlihatkan ketrampilan tertentu yang menunjukkan hasil belajar, atau dalam menggunakan kemampuan psikologis untuk memecahkan suatu persoalan (Djemari Mardapi, 2004). Tes yang berkualitas berkaitan dengan soal yang berkualitas. Soal yang berkualitas memenuhi kriteria sebagai alat ukur yang baik. Kriteria alat ukur yang baik dapat ditinjau dari teori tes klasik (*Classical Test Theory*: CTT) dan teori tes modern atau teori respons item *(Item Responsse Theory*: IRT). Parameter yang diukur pada CTT yaitu reliabilitas tes, indeks daya beda item (*discrimination index*), indeks kesukaran item (*difficulty index*), validitas isi (*content validity*), validitas konstruk (*construct validity*), dan validitas berdasar kriteria (*criterion related validity*). Parameter yang diukur pada IRT yaitu kemampuan (*ability)* dan parameter item yang terdiri atas indeks daya beda item (*discrimination index*), indeks kesukaran item (*difficulty index)*, dan terkaan (*guessing*). Salah satu kriteria alat ukur yang baik memilik terkaan (*guessing*) yang relatif kecil.

Teori tes klasik digunakan secara luas pada pengukuran pendidikan di Indonesia, walaupun memiliki keterbatasan. Salah satu keterbatasan dari teori tes klasik yaitu, ketika tes yang sama diberikan pada siswa yang berbeda, dan hasilnya dinyatakan dengan skor total. Tingkat kemampuan siswa tidak dapat dibandingkan berdasarkan skor total yang diperoleh siswa tersebut, karena skor total yang diperoleh siswa tidak menunjukkan tingkat kesulitan tes yang dikerjakan. Seorang siswa yang kemampuannya lebih rendah mungkin mendapatkan skor lebih tinggi pada suatu tes yang mudah dibandingkan dengan siswa lain yang lebih tinggi kemampuannya tetapi mendapatkan tes yang lebih sulit. Keadaan tersebut menggambarkan bahwa skor total tidak dapat dibandingkan, karena kedua siswa tersebut mengerjakan item yang mungkin memiliki tingkat kesulitan berbeda, walaupun skor total yang dicapainya sama.



Gambar 1. Kemampuan testee pada CTT

Pada kegiatan pembelajaran di kelas, guru menggunakan teori tes klasik untuk mengidentifikasi karakteristik soal yang digunakan, serta mengidentifikasi tingkat kemampuan siswa berdasarkan skor total yang diperoleh siswanya. Pada teori respons item (IRT), siswa memiliki sifat *invariance* terhadap tingkat kesulitan item dan sebaliknya. Tingkat kemampuan siswa tidak tergantung pada tingkat kesulitan item, dan tingkat kesulitan item tidak tergantung dari siswa yang mengerjakannya. Penggunaan skor dari dua tes yang disetarakan menuntut kedua tes memiliki tingkat kesulitan setara. Proses statistik yang digunakan untuk menyetarakan skor kedua tes tersebut, disebut penyetaraan atau *equating*. Hambleton & Swaminathan (1985: 123) menyatakan bahwa penyetaraan merupakan prosedur statistik untuk menetapkan hubungan antara skor dari dua tes atau lebih. Kolen & Brenan (2004: 2) menyatakan bahwa proses penyetaraan dapat dilakukan untuk menyetarakan dua tes atau lebih dengan materi dan tingkat kesulitan yang setara. Penggunaan skala kemampuan yang sama pada penyetaraan skor tes memiliki keuntungan antara lain memungkinkan dilakukannya evaluasi terhadap hasil tes, dapat mengembangkan tes yang setara dan skornya dapat dipertukarkan, memungkinkan terjaminnya keamanan tes, serta dapat dikembangkannya bank soal.

Hambleton & Swaminathan (1985) menyatakan bahwa prosedur penyetaraan skor tes atau *test equating* ada dua cara, yaitu cara penyetaraan *vertical* dan cara penyetaraan h*orizontal.* Cara penyetaraan *vertical* adalah suatu usaha untuk menyetarakan skor tes pada dua tes atau lebih yang dirancang berbeda tingkat kesulitan tetapi mengukur isi dan jenis kemampuan yang sama. Penyetaraan *vertical* dirancang untuk kontinuitas tes. Kontinuitas tes diartikan sebagai keberlanjutan tes yang digunakan untuk mengukur perkembangan atau perubahan tingkat kemampuan siswa. Cara penyetaraan *horizontal* dilakukan pada tes paralel yang memiliki kesamaan isi dan tingkat kesulitan, kemudian diberikan pada kelompok siswa yang memiliki tingkat kemampuan setara.

Menurut Kolen & Brenan (2004: 13) untuk meminimalisir ketidakakuratan hasil penyetaraan, diperlukan rancangan penyetaraan. Rancangan penyetaraan beragam, ada rancangan grup tunggal (*Single Group Design:* SG), rancangan grup ekivalen (*Equivalent Group Design:* EG), rancangan grup yang diseimbangkan (*Counter Balanced Design:* CB), rancangan pengait (*Anchor Test Design: AT)* atau (*Non Equivalent Anchor Test Design*: NEAT). Rancangan tersebut memiliki karakteristik berbeda serta memiliki kelebihan dan kekurangan sendiri.

Penyetaraan skor tes menggunakan teori respons item (*Item Responsse Theory*: IRT) merupakan model penyetaraan yang lebih *representatif* dibandingkan dengan model penyetaraan menggunakan teori tes klasik (*Classical Test Theory*: CTT). Teori respons item memiliki sifat *invariance* pada parameternya. Parameter kemampuan (*ability*) siswa *invariance* terhadap parameter tes dan sebaliknya. Oleh sebab itu, tes yang dikerjakan siswa tetap pada skala yang sama selama fungsi informasi tes tinggi. Model teori respons item ada dua, yaitu model teori respons item untuk data *dichotomous* (dikotomos) dan model teori respons item untuk data *polytomous* (politomos). Data dikotomos mempunyai dua kemungkinan jawaban yaitu benar atau berhasil (antara lain dinyatakan dengan skor 1), dan salah atau gagal (antara lain dinyatakan dengan skor 0). Data politomos mempunyai kemungkinan jawaban lebih dari dua (Embretson & Reise, 2000: 14).

Tujuan dari penelitian ini yaitu 1) menemukan akurasi estimasi parameter item pada *test equating* menggunakan metode *Item Characteristic*

Curve (ICC) yang ditunjukkan oleh nilai RMSD dari parameter item sebelum dan sesudah penyetaraan, 2) Menemukan sensitivitas metode *linear* yang terdiri atas *Tucker - Levine score method* dan *Levine true score method applied to observed scores method* serta metode *equipercentile* yang terdiri atas *Braun-Holland linear method* dan *chained equipercentile equating method* yang ditunjukkan oleh nilai RMSE sebelum dan setelah penyetaraan.

Metode Penelitian

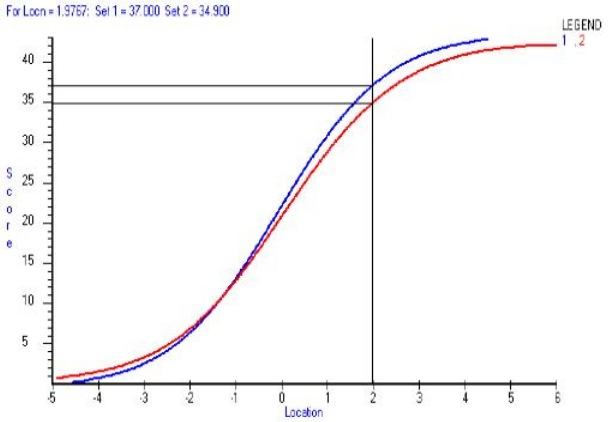
Rancangan penyetaraan yang digunakan yaitu rancangan pengait (*Anchor Test Design: AT)* atau (*Non Equivalent Anchor Test Design*: NEAT). NEAT desain digambarkan sebagai berikut:

Tabel 1.
Anchor Design (NEAT)

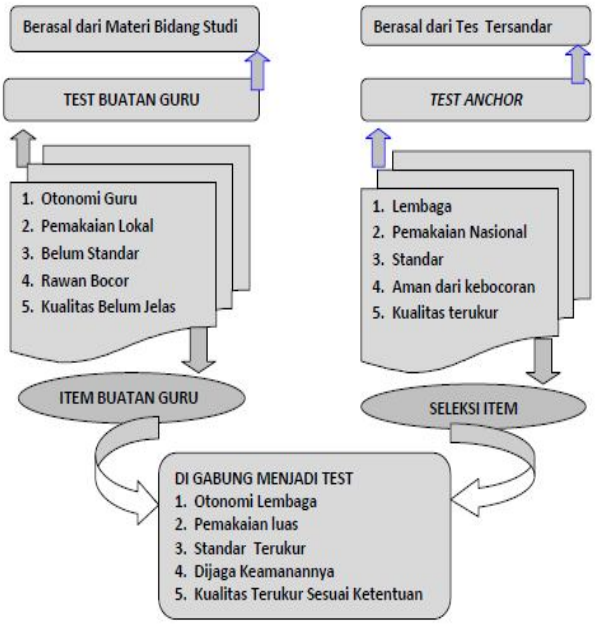
Population	Sampel	X	Z	Y
P	1	√	√	-
Q	2	-	√	√

Metode penyetaraan untuk data dikotomos pada *nonequivalent groups design* dapat dilakukan dengan cara, *item responsse theory* (IRT) *method, linear method* dan *equipercentile method*. *Test equating* menggunakan IRT dilakukan melalui 3 langkah. Pertama, parameter item dihitung menggunakan program komputer. Kedua, estimasi skala parameter menggunakan transformasi linear berdasarkan IRT. Ketiga, mengklassifikasikan parameter item bedasarkan hasil analisis program olah data komputer. Konsep dan pemahaman IRT memberikan dasar pada *test equating* untuk data dikotomos.

P_j ditulis sebagai fungsi dari variabel θ. Perbedaan model IRT dengan lainnya yaitu asumsi dari bentuk fungsional yang digambarkan dengan *Item Characteristic Curve* (ICC) atau kurva karakteristik item, seperti Gambar 1.



Gambar 1. *Item Characteristic Curve* (ICC) *Model*



Gambar 2. Proses Penyetaraan Menggunakan NEAT desain

Desain pengumpulan data menggunakan *anchor design* atau *Non Equivalent Anchor Test design* (NEAT), dan metode analisis penyetaraan yang digunakan yaitu metode *Item Characteristic Curva* (ICC), metode *linear* dan metode *equipercentile.* Metode ICC menggambarkan plot setiap item dalam bentuk kurva. Metode *linear* terdiri atas *Tucker-Levine observed score methods* dan *Levine true score method applied to observed scores.* Metode *equipercentile* terdiri atas *Braun – Holland linear methods* dan *chained equipercentile equating methods*. Penelitian yang dilakukan menggunakan data skor amatan dan skor sebenarnya.

Hasil Penelitian

Berdasarkan rumusan masalah pada BAB I yaitu bagaimana akurasi estimasi parameter item pada *test equating* dengan metode *Item Characteristic Curve* (ICC) yang ditunjukkan oleh nilai RMSD dari parameter item sebelum dan sesudah penyetaraan, dan bagaimana sensitivitas *linear method* yang terdiri atas *Tucker - Levine score method* dan *Levine true score method applied to observed scores* serta *equipercentile method* yang terdiri atas *Braun-Holland linear method* dan *chained equipercentile equating method* yang ditunjukkan oleh nilai RMSE sebelum dan setelah penyetaraan.

Tabel 2. Statistik Deskriptif Estimasi Parameter Item Pra dan Post Penyetaraan Tes A, B dan C

Parameter Item	Besaran yang diukur	Tes A	Tes B	Tes C	Hasil Penyetaraan
Slope (a)	MEAN	0.663	0.916	0.811	0.746
	SD	0.261	0.212	0.168	0.282
	N	55	55	50	140
Threshold (b)	MEAN	-0.058	0.318	-0.198	0.161
	SD	1.108	0.731	0.983	1.195
	N	55	55	50	140
RMSD CHI-SQUARE	0.0776				

Tabel 3. Akurasi Estimasi Parameter Item pada *Test equating* Menggunakan Metode ICC Tes A, B, dan C

ICC	$b_B = 1.88\, b_A - 0.5$ $a_B = \frac{a_A}{1{,}88}$	$b_C = 1.45\, b_B - 0.23$ $a_C = \frac{a_B}{1{,}45}$	$b_C = 2.72\, b_A - 0.02$ $a_C = \frac{a_A}{2{,}72}$
A	**RMSD**	a_A	0.0712
		b_A	0.0137
B	**RMSD**	a_B	0.0378
		b_B	0.0140
C	**RMSD**	a_C	0.0371
		b_C	0.0140
A + B + C	**RMSD**	**a**	**0.0487**
		b	**0.0139**

Tabel 4. Hasil Berbagai Metode Penyetaraan Tes A, B dan C

Levine – true score method	*Tucker - Levine* Skor amatan	*Braun - Holland linear*	*Chained Equipercentile Equating*
$\gamma_1 = 112.5958$	$l_{Y_s}(A) = 0.9766X - 5.5578$	$\sigma_2^2(A) = \frac{\sigma^2}{n}(A) = 0.1202$	$e_Y(B) = 28.4495$
$\gamma_2 = 46.17211$	$l_{Y_s}(B) = 0.5980X + 0.1552$	$\sigma_2^2(B) = \frac{\sigma^2}{n}(B) = 0.1042$	$e_Y(C) = 11.2892$
$\gamma_1/\gamma_2 = 2.4387$	$l_{Y_s}(C) = 1.7531X - 0.1839$	$\sigma_2^2(C) = \frac{\sigma^2}{n}(C) = 0.1008$	$e_Y(A) = 11.2892$
$l_Y(x) = 2.4387X - 4.6035$			

Tabel 5. Sensitivitas (RMSE) Berbagai Metode Penyetaraan Tes A, B dan C

KLP	*Metode Linear*		*Metode Equipercentile*	
	Levine	*T - Levine*	*B - Holland*	*Chained*
A	0.0023	0.0023	0.0089	0.0111
B	0.045	0.0127	0.0339	0.0444
C	0.000	0.0150	0.0267	0.0165
Rerata RMSE	0.01577	0.01	0.0232	0.0279
	0.0129		0.0255	

Simpulan

Simpulan yang diperoleh, akurasi estimasi parameter item pada *test equating* menggunakan metode ICC berdasarkan nilai *root mean square difference* (RMSD) dari estimasi parameter item, pada kedua parameter item relatif baik (RMSD ≤ 0.1 untuk indeks daya beda item, dan RMSD ≤ 0.1 untuk indeks kesulitan item). Secara keseluruhan, sensitivitas berbagai metode penyetaraan berdasarkan hasil hitung *root mean square error* (RMSE ≤ 0,1), menunjukkan sensitivitas dari berbagai metode penyetaraan relatif tinggi. Penyetaraan menggunakan metode *linear* memiliki sensitivitas yang lebih tinggi dibandingkan dengan metode *equipercentile*. Sensitivitas metode penyetaraan yang paling tinggi yaitu *Tucker – Levine method.* Urutan sensitivitas metode penyetaraan dari paling tinggi sampai paling rendah yaitu *Tucker – Levine method, Levine method*, *Braun - Holland linear method* dan *Chained Equipercentile Equating method.*

Implikasi Berbagai Metode *Test Equating* Pemilihan metode yang paling cocok untuk *test equating* tergantung dari terpenuhinya asumsi setiap metode *test equating*. Oleh sebab itu pengujian asumsi sebelum dilakukan *equating* menjadi suatu kegiatan yang tidak dapat ditinggalkan. Terpenuhinya asumsi memberi kontribusi yang optimal pada hasil *test equating.* Hal ini didukung oleh penelitian yang dilakukan oleh Alina (2005) tentang pendekatan penyetaraan *linear* untuk *non – equivalent groups design* (NEAT)

pada tiga metode penyetaraan linear dalam NEAT *design*, yaitu Tucker, Levine skor amatan, *chain* dan pengembangan suatu *common* parameterisasi. Dia menyatakan bahwa setiap metode penyetaraan adalah suatu kasus khusus pada fungsi penyetaraan *linear* dalam NEAT *design.* Pemilihan metode *linear* untuk *test equating* merupakan pilihan yang tepat jika asumsi dipenuhi. Jika asumsi tidak dipenuhi dapat dipilih metode lain, misalnya metode *Braun - Holland linear* atau Metode *Chained equipercentile equating*, sesuai dengan asumsi yang dituntut kedua metode tersebut.

Perencanaan tes untuk skala besar perlu dikondisikan. Hal ini member kesempatan pada pengajar bidang studi bekerja dalam *team work*. Merencanakan dan menyusun instrumen sendiri berbeda dengan merencanakan dan menyusun instrumen dalam *team work.* Instrumen yang dikerjakan dalam *team work* memiliki kualitas yang relatif lebih baik dibandingkan dengan instrumen tes yang direncanakan dan disusun sendiri. Tes berskala besar memudahkan pengembangan bank item, keberadaan bank item diperlukan untuk meningkatkan kualitas tes.

Pada penelitian ini, *anchor test* yang dipilih bersifat *external*. *External anchor test* dirancang dan disusun untuk mengukur kemampuan berfikir tingkat tinggi atau *high other thinking* (HOT). Item yang digunakan untuk *anchor* pada *equating* dipilih secara acak dari tes HOT yang telah teruji. Implikasi dari penyusunan tes yang terencana, memberi kontribusi pada peningkatan kualitas tes pada khususnya dan kualitas pendidikan pada umumnya.

Berdasarkan hasil olah data, pembahasan dan simpulan yang telah dikemukakan, menyarankan kepada peneliti yang berkecimpung dalam penelitian psikometri, Penelitian metode *test equating* untuk data dikotomos dengan *external anchor test* dapat dikembangkan untuk mata pelajaran selain Ilmu Pengetahuan Alam (IPA). Makin banyak penelitian yang dilakukan, makin memperkaya instrumen pengukuran yang berkualitas. Instrumen yang berkualitas memberi kontribusi pada peningkatan kualitas pendidikan pada umumnya. Metode *linear* memiliki sensitivitas yang lebih baik dibandingkan metode *equipercentile* jika asumsi dipenuhi. Oleh sebab itu metode *linear* dapat dipilih untuk *test equating.* Metode *test equating* perlu disosialisasikan pada lembaga yang berkecimpung dalam pengujian sebagai

dasar penyusunan bank soal dan pemetaan kualitas pendidikan. Selain itu *test equating* menggunakan *external anchor test*, mampu mengeliminasi kelemahan tes buatan guru dan mengatasi kekurangan dari tes yang terstandar.

Daftar Pustaka

Allen, M.J. & Yen,W.M. (1979). *Introduction to measurement theory.* Monterey, CA: Brooks/ Cole Publishing Company.

Avi. (2007). An NCME instructional module on quality control procedures in the scoring, equating, and reporting of test scores. *Educational Measurement*, Vol 26, Edisi 1, pp.36 - 43.

Brennan, R.L. (2006), *Educational measurement*. Iowa City*:* United State of America: American Council on Education and Praeger Publisher.

Brennan, R.L. & Kolen, M.J. (2004), *Test equating, scaling, and linking*. Iowa City*:* United State of America: American Council on Education and Springer Publisher.

Crocker & Algina. (1986), *Introduction to classical and modern test theory*. New York: United State of America: CBS College Publishing.

Djemari Mardapi. (2004). *Penyusunan tes hasil belajar*. Yogyakarta: Program Pascasarjana Universitas Negeri Yogyakarta.

Djemari Mardapi. (2008). *Teknik penyusunan instrumen tes dan non tes.* Yogyakarta: Mitra Cendikia Press.

Dikdasmen Dikbud (1999). *Pengelolaan pengujian bagi guru mata pelajaran*. Jakarta: Departemen Pendidikan dan Kebudayaan Direktorat Jenderal Pendidikan Dasar dan Menengah Direktorat Pendidikan Menengah Umum.

Embretson & Reise. (2000), *Item response theory for psychologists*. London : Lawrence Erlbaum associates publishers.

Ebel R.L. & Friesbie. D.A (1986), *Essentials of educational measurement*. New Jersey: Prentice – Hall. Inc

Ekohariadi. (2009). *Perkembangan kemampuan sains siswa Indonesia berusia 15 tahun*

berdasarkan data studi PISA. Makalah Seminar mutu Pendidikan Dasar dan Menengah Hasil Penelitian Puspendik. Jakarta. Badan Penelitian dan Pengembangan Departemen Pendidikan Nasional.

Freund's, J. E. (2004). *Mathematical statistics with applications,* Canada: Pearson Education.Inc.

Gary, S. (2005). Acuracy of random groups equating with very small samp*les. Journal of Education Measurement*. Vol 42 number 4 winter 2005. p. 309 – 330.

Hambleton, R. K., Swaminathan, H., & Rogers, H.J. (1991). *Fundamental of item response theory.* Newbury Park, CA: Sage Publication. New Inc.

Hambleton, R.K. & Swaminathan, H. (1985). *Item response theory principles and applications.* Boston, MA: Kluwer Inc.

Kolen, M.J. & Brenan, R.L. (1995). *Test equating.* Iowa City : Springer. Kolen, M.J. & Brenan,R.L. (2004). *Test equiting, scaling, and linking..*Iowa City: Springer.

Kolen, M.J. (2004). Linking assessment : concept and history. *Applied Psychological Measurement* , Vol. 28, No 4, pp . 219 – 226.