PERBANDINGAN RELIABILITAS TES HASIL BELAJAR MATEMATIKA SMA BERDASARKAN TEKNIK PENSKORAN DAN UKURAN SAMPEL

Dwi Putri Musdansi[1], Nahadi[2]
[1]Universitas Islam Kuantan Singingi
[2]Universitas Pendidikan Indonesia
dwiputrimusdansi@gmail.com

Abstract

The purpose of this research was to know the comparison of the scoring techniques number right score and reward score, number right score and punishment, than sample size 30 dan 40 on reliability of the mathematics achievement test. The instrument was used multiple choice form test for High School. Sampling teqniques was cluster random sampling. This research used a static group comparison design. Analyze by using t-test. The research result were (1) there was a the difference between scoring techniques number right score and reward score on reliability of multiple-choice on Mathematics achievement tests, (2) there was a the difference between scoring techniques number right score and punishment score on reliability of multiple-choice on Mathematics achievement tests, (3) for the group of reward score, there was not the difference on reliability of multiple-choice on Mathematics achievement tests between sample size 30 and 40, (4) for the group of punishment score, there was not the difference on reliability of multiple-choice on Mathematics achievement tests between sample size 30 and 40.

Keywords: *Scoring Technique, Sample Size, Reliability*

1. PENDAHULUAN

Tujuan belajar adalah untuk mengadakan perubahan didalam diri seperti mengubah kebiasaan dari yang buruk menjadi baik, mengubah sikap dari yang negatif menjadi yang positif (Dalyono, 2010:49). Dengan belajar seseorang dapat menambah pengetahuan dalam berbagai bidang ilmu, dapat berfikir kritis, kreatif dan lain sebagainya. Belajar memiliki tiga prinsip, yaitu *pertama,* prinsip belajar sebagai perubahan perilaku. *Kedua,* belajar sebagai proses. *Ketiga*, belajar sebagai bentuk pengalaman (Suprijono, 2009:4)

Dalam proses belajarn guru dianggap menjadi unsur paling menentukan keberhasilan peserta didik, maka sudah sewajarnyalah dalam proses pembelajaran guru selalu melakukan berbagai evaluasi, pengukuran, serta penilaian sebagai umpan balik dalam upaya memperbaiki proses belajar mengajar, mengukur sejauhmana keberhasilan siswa, maupun sebagai informasi yang berharga dalam mengambil keputusan. Informasi tersebut bisa didapat dengan melakukan prosedur pengukuran, karena hanya dengan melakukan pengukuran dapat dikuantifikasikan secara benar, yaitu melalui pengamatan indikator-indikatornya yang operasional. Oleh karena itu, untuk mendapatkan semua informasi hasil ukur yang tidak menyesatkan, tentu mebutuhkan alat ukur yang berkualitas tinggi sehingga informasi yang diperoleh dapat dipercaya (Ali, 2011).

Apabila alat ukur yang dibuat memberikan informasi yang tidak akurat digunakan dalam pengambilan keputusan, maka dapat dipastikan kesimpulan atau keputusan tersebut pastilah juga merupakan keputusan yang tidak tepat. Keputusan yang tidak tepat memang tidak langsung dapat dirasakan akibatnya, namun dapat menimbulkan hal buruk yang tidak dikehendaki. Sebagai contoh seseorang dapat ditolak oleh perguruan tinggi berdasarkan tes masuk yang diselenggarakan, padahal mahasiswa tersebut sangat potensial. Dari hal tersebut terlihat bahwa dengan menggunakan tes yang tidak mampu menghasilkan data yang reliabel dan valid, bukan saja calon mahasiswa yang bersangkutan yang dirugikan tetapi juga Universitas yang menolak boleh jadi kehilangan mahasiswa yang sangat potensial. Contoh tersebut memberi pelajaran bahwa betapa besar dampak yang ditimbulkan jika instrumen yang digunakan tidak reliabel dan valid jika digunakan dalam mengambil keputusan.

Oleh karenanya, para ahli psikometrika telah menetapkan beberapa kriteria penting bagi setiap alat ukur psikologi untuk dapat dinyatakan sebagai alat ukur yang baik, yaitu mampu menghasilkan data dan memberikan informasi yang akurat. “Kriteria yang dimaksud adalah valid, reliabel, objektif, standar, ekonomis dan praktis” (Azwar, 2012:2). Naga (2013) mengatakan alat ukur yang baik memberikan hasil yang konstan bila digunakan berulang-ulang, asalkan kemampuan yang diukur tidak berubah. Hal tersebut berkaitan dengan reliabilitas tes tersebut. Makin tinggi reliabilitasnya maka akan tinggi pula tingkat kepercayaan skor amatannya. Skor amatan akan mendekati skor tulennya jika komponen keliru dalam pengukuran semakin kecil.

Ada beberapa hal yang dapat mempengaruhi reliabilitas tes antara lain perbedaan kondisi individu, perbedaan kondisi tes, variasi pengadministrasian tes, serta kesalahan dan perbedaan pemberian skor (Busnawir,2006). Disisi lain juga perlu dipertimbangkan faktor panjang tes, kecepatan, homogenitas kemampuan siswa, tingkat kesukaran butir yang dapat mempengaruhi reliabilitas tes (Busnawir, 2006). Menurut Harun (2007:131) reliabilitas juga dipengaruhi oleh cara penyajian tes, suasana hati, sikap subjek terhadap tes, motivasi dan kondisi subjek, keadaan ruangan pengujian, cara memberikan tes, dan sebagainya. Seperti juga dikemukakan Surapranata (2009:87) bahwa salah satu dari enam faktor penyebab terjadinya perbedaan skor (interkonsistensi) adalah faktor yang tidak pernah diperhitungkan seperti menerka dan melihat soal yang dilihat sebelumnya.

Sebagaimana diketahui bahwa untuk menetapkan nilai yang berkaitan dengan kinerja dan hasil karya siswa dilakukan melalui evaluasi. Cara evaluasi yang umum

dilakukan oleh pelaku pendidikan adalah dengan menggunakan tes. Tes dianggap sebagai alat ukur yang paling praktis untuk mengetahui kemampuan siswa. Seperti yang dikemukan oleh Anastasi (2007) bahwa tes pada dasarnya adalah alat ukur yang objektif dan dibakukan atas sampel tertentu. Mengutip pendapat Frenderick G. Brown dari buku Azwar: Frenderick (Azwar, 2011:3) mengemukakan bahwa tes adalah "prosedur yang sistematik guna mengukur sampel perilaku seseorang. Tes dapat disusun dalam berbagai bentuk dan tipe sesuai dengan tujuan dan maksud penyusunan tes".

Bentuk tes yang sampai saat ini paling banyak digunakan dalam dunia pendidikan adalah tes bentuk pilihan ganda. Tes bentuk pilihan ganda banyak digunakan karena memiliki kekuatan tersendiri. Zainul (1993:62-64) mengatakan bahwa *pertama*, butir soal tipe pilihan ganda dapat dikonstruksi dan digunakan untuk mengukur segala level tujuan intruksional, mulai dari yang sederhana sampai yang paling kompleks. *Kedua*, karakteristik dari butir pilihan ganda hanya menuntut waktu kerja peserta tes sangat minimal sehingga dapat menggunakan jumlah butir soal relatif banyak yang berimplikasi pada penarikan sampel pokok bahasan yang akan diujikan dapat lebih luas. *Ketiga*, penskoran hasil kerja peserta dapat dikerjakan secara objektif. *Keempat*, tipe butir soal dapat dikonstruksi sehingga menuntut kamampuan peserta tes untuk membedakan berbagai tingkatan kebenaran sekaligus. *Kelima*, jumlah *option* yang dapat disediakan lebih dari dua. *Keenam*, tipe butir soal pilihan ganda memungkinkan dilakukan analisis butir soal secara baik. *Ketujuh*, tingkat kesukaran butir soal dapat dikendalikan dengan mengubah tingkat homogenitas alternatif jawaban. *Kedelapan*, informasi yang diberikan lebih kaya.

Selain kekuatan-kekuatan tersebut, tes pilihan ganda juga memiliki keterbatasan yaitu sukar dikonstruksi, ada kecenderungan pembuat tes mengkonstruksi butir soal tipe ini hanya mengukur ranah ingatan, atau aspek yang paling rendah yaitu ranah kognitif saja. Selain itu setiap jawaban butir tipe objektif memiliki peluang peserta tes untuk melakukan terkaan atau tebakan. Tingkat terkaan adalah satu per jumlah *option* dalam butir soal itu. Terkaan pada soal pilihan ganda sebenarnya dikarenakan tidak adanya antisipasi agar pengikut tes tidak melakukan tebakan. Antisipasi ini berkaitan dengan teknik penskoran. Teknik penskoran yang saat ini umum digunakan adalah penskoran menjumlahkan jawaban betul atau disebut juga *number right score*. Dengan teknik ini, bisa saja soal sukar direspon benar oleh siswa yang tidak memiliki kecukupan pengetahuan, dan dijawab salah oleh yang memiliki kecukupan pengetahuan. Jawaban benar yang diberikan oleh peserta tes yang

memiliki kemampuan rendah dan jawaban salah oleh kemampuan tinggi tentu berakibat pada reliabilitas yang dihasilkan. Apalagi dalam pelajaran Matematika yang oleh kebanyakan siswa dijadikan momok tersendiri, sehingga besar kemungkinan siswa untuk melakukan terkaan terhadap tes yang diberikan.

Karena adanya faktor terkaan yang ditimbulkan oleh tes bentuk pilihan ganda terutama dalam pelajaran Matematika, tentu hal tersebut memberikan masalah, yaitu hasil tes yang diperoleh peserta tes tidak menggambarkan kemampuan peserta tes yang sesungguhnya sehingga timbul ketidakpercayaan orang terhadap hasil tersebut. Selain itu juga tidak akan terdeteksi apakah siswa memahami tentang soal yang ada atau tidak. Oleh karenanya, untuk meminimalisir terkaan siswa pada tes pilihan ganda maka dibuatlah teknik penskoran untuk mengurangi tebakan. Teknik penskoran tersebut adalah penskoran pinalti. " Pemberian skor dengan pinalti (hukuman) dalam tes prestasi dikenal dengan istilah *Correction for guessing*"(Azwar, 2011 :114). Metoda penskoran dengan pinalti ini diyakini dapat menigkatkan reliabilitas dan validitas tes. Hal ini dikarenakan skor yang terkoreksi sudah memiliki tingkat estimasi yang lebih baik sebab meliputi pengukuran karakteristik adanya terkaan (Croker dan Algina: 401). Selain peskoran dengan hukuman, upaya lain untuk meningkatkan reliabilitas dengan mengurangi efek terkaan adalah adanya tambahan nilai terhadap butir soal yang diragukan dengan cara mengosongkan jawaban tersebut, dikenal dengan *reward score.* Kedua teknik *Correction for guessing* tersebut belum atau tidak pernah digunakan untuk oleh para guru disekolah sebagai antisipasi agar siswa tidak melakukan tebakan. Karena guru belum tahu persis besar pengurangannya maupun kebermanfaatan mengenai penggunaan teknik ini.

Selain itu, mengutip pendapat Nikto (1996:72-73) mengatakan bahwa "hal yang perlu diperhatikan dalam estimasi kestabilan koefisien reliabilitas adalah ukuran sampel, karena indek reliabilitas dihitung berdasarkan sampel responden sehingga akan berfluktuasi berdasarkan besar sampel yang ditarik dari suatu populasi". Namun seberapa besar perbedaan fluktuasi itu belum dapat ditentukan secara pasti. Azwar menyebutkan bahwa semakin heterogen suatu kelompok maka reliabilitas yang dihasilkan semakin tinggi. Sebaliknya, semakin homogen suatu kelompok maka semakin kecil reliabilitasnya. Heterogenitas ataupun homogenitas kelompok diperlihatkan oleh besar kecilnya varians distribusi skor subjek pada variabel yang diungkap oleh tes yang bersangkutan.

Dari uraian diatas dikemukakan bahwa jika menulis tes dalam bentuk tes pilihan ganda maka akan berpeluang untuk melakukan

tebakan. Ada atau tidaknya tebakan tergantung pada intervensi yang diberikan pada tes, yaitu berupa petunjuk soal pada jenis teknik penskoran. Disisi lain reliabilitas juga dipengaruhi oleh ukuran sampel yang ditarik dari populasi, karena reliabilitas dihitung berdasarkan ukuran responden. Artinya reliabilitas akan berfluktuasi sesuai dengan jumlah sampel yang ditarik. Hal itu disebabkan karena adanya perbedaan sebaran skor yang menyebabkan jika semakin besar ukuran sampel maka akan diperoleh variasi skor yang lebih beragam daripada ukuran sampel yang kecil. Oleh sebab itu melalui penelitian ini, peneliti ingin mengkaji lebih jauh mengenai:

1. Perbedaan reliabilitas tes pilihan ganda hasil belajar Matematika antara teknik penskoran *number-right score* dan *reward score*
2. Perbedaan reliabilitas tes pilihan ganda hasil belajar Matematika antara teknik penskoran *number-right score* dan *punishment score*
3. Perbedaan reliabilitas tes pilihan ganda hasil belajar Matematika kelompok *reward* dengan ukuran sampel 30 dan 40
4. Perbedaan reliabilitas tes pilihan ganda hasil belajar Matematika kelompok *punishment* dengan ukuran sampel 30 dan 40

2. Metode

Penelitian ini merupakan perbandingan reliabilitas tes hasil belajar matematika berdasar metode penskoran *number-right score* dan metode penskoran *correction for guessing*. Penelitian ini termasuk kedalam penelitian kuantitatif dengan menggunakan metode kuasi-eksperimental (eksperimen semu). Dikatakan kuasi eksperimental karena peneliti tidak melakukan *random assigment* saat melakukan penskoran menggunakan teknik penskoran *number-right score*, *punishment score* dan *reward score.*

Desain penelitian yang digunakan adalah desain perbandingan kelompok statis (*the static group comparison design*) yaitu *design* yang dapat digunakan untuk membandingkan dua atau tiga kelompok penelitian. Satu kelompok kontrol reliabilitas skor hasil tesnya menggunakan metode penskoran *number-right score*. Sedangkan dua kelompok eksperimen lain reliabilitas skor hasil tesnya menggunakan metode penskoran *punishment score* dan *reward score.*

Tabel 1. Desain penelitian

Teknik Penskoran		
Number-right score	**Correction for guessing**	
	Punishment score	***reward score***
$X_{NR} = \sum_{i=1}^{n} x_i$ $r_{NR1}, r_{NR2},\ldots, r_{NR30}$ μ_{NR}	$X_{cp} = R - \frac{W}{k-1}$ $r_{P1}, r_{P2},\ldots, r_{p30}$ μ_P	$X_{cr} = R + \frac{O}{k}$ $r_{R1}, r_{R2},\ldots, r_{R30}$ μ_R

Peneliti mengumpulkan data dengan instrumen tes hasil belajar Matematika kelas X SMA semester 2 materi Geometri dan Limit Fungsi. Tes terdiri dari 25 butir soal pilihan ganda yang telah diuji tingkat kesukaran, daya beda, validitas serta reliabilitasnya. Penarikan sampel data dilakukan dengan teknik pengembalian (*random sampling with replacement).* Teknik *random* dilakukan dengan bantuan program *IBM SPSS Statistics 20*, yaitu teknik pencuplikan sederhana. Penghitungan koefisien reliabilitas itu dilakukan sebanyak 30 kali, sehingga akan diperoleh 30 koefisien reliabilitas pada setiap teknik penskoran.

3. Hasil dan Pembahasan

1. ada perbedaan reliabilitas antara teknik penskoran *number-right score* dan *reward score*

Teknik penskoran *reward* adalah penskoran dengan cara menjumlahkan skor jawaban benar dan menambahkan skor jawaban yang dikosongkan oleh setiap siswa pada masing-masing butir soal. Butir jawaban benar diberi skor 1, jawaban salah diberi skor 0, dan jawaban yang dikosongkan diberi skor 0,2. Jadi maksimum skor yang diperoleh setiap siswa adalah 25. Sedangkan *number-right* merupakan cara menjumlahkan jawaban benar saja. Jawaban benar diberi skor 1, jawaban salah dan dikosongkan diberi skor 0. Dalam penelitian ini, siswa yang dikenakan teknik penskoran *reward score* sekaligus *number-right score* berjumlah 153 siswa.

Tabel.2 Hasil Uji hipotesis 1

Paired Samples Test

		Paired Differences					T	df	Sig. (2-tailed)
		Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference				
					Lower	Upper			
Pair 1	NR - R	,068367	,028810	,005260	,057609	,079125	12,997	29	,000

Hasil analisis diperoleh hasil t= 12,997, sig= 0,000. Dengan menggunakan taraf signifikansi 5% maka $Pvalue < 0{,}05$, dengan demikian disimpulkan bahwa H_0 ditolak. Ini mengindikasikan bahwa ada perbedaan *significant* reliabilitas antara teknik penskoran *number right score* dan *reward score*. Jika dilihat dari nilai rata-rata reliabilitas, terlihat bahwa *mean* reliabilitas *number right score* lebih besar dari *mean* reliabilitas *reward score*, yaitu 0,687 > 0,619.

Sebelumnya dibahas, dengan teknik penskoran *number right score*, bisa saja soal sukar direspon benar oleh siswa yang tidak memiliki kecukupan pengetahuan, dan dijawab salah oleh yang memiliki kecukupan pengetahuan. Jawaban benar yang diberikan oleh peserta tes yang memiliki kemampuan rendah dan jawaban salah oleh kemampuan tinggi tentu berakibat pada reliabilitas yang dihasilkan, yaitu akan diperoleh reliabilitas yang rendah. Karena peserta tes yang tidak memiliki kecukupan pengetahuan dapat menjawab butir soal yang sukar dengan benar, dapat diindikasikan peserta tersebut melakukan penebakan. Sehingga perlu alternatif teknik lain untuk mengurangi penebakan itu. Teknik penskoran *reward* merupakan salah satu teknik *Corection for geussing*. Teknik ini bertujuan untuk meminimalisir *guessing* agar reliabilitas lebih baik. Tapi analisis menunjukan bahwa rata-rata reliabilitas dengan teknik *number right score* lebih tinggi daripada *reward score*. Kesimpulan ini juga sejalan dengan pendapat Venny (2009) pada penelitiannya yang berjudul perbandingan daya beda dan reliabilitas tes pilihan ganda berdasarkan model penskoran siswa menyimpulkan bahwa model penskoran komposit memiliki reliabilitas lebih besar dari model penelitian pinalti penskoran kompensasi

Tapi dalam penelitian ini bukan berarti teknik *number right score* lebih baik daripada *reward score*. Perlu diingat, bahwa dalam penelitian ini perolehan skor pada *number right score* merupakan hasil respon dari instrumen *reward* dengan memberi skor 1 pada jawaban benar dan skor 0 pada respon lainnya, sehingga skor yang dihasilkan pada teknik *number right score* benar-benar skor yang sesuai dengan

kemampuan pengikut tes. Artinya pengikut tes yang memiliki kecukupan kemampuan akan merespon tes dengan benar, sebaliknya pengikut tes yang tidak memiliki kecukupan pengetahuan akan merespon tes dengan salah.

Selain itu, pada teknik *reward* jawaban yang salah (yang diasumsikan sebagai skor tebakan) akan diberi skor 0. Sehingga, Bila dianalogikan reliabilitas yang dihasilkan oleh *number-right* skor sebagai reliabilitas *reward score,* dengan skor 0 merupakan hasil tebakan, maka tebakan pada *reward score* akan meningkatkan reliabilitas. Dengan demikian teknik penskoran *reward score* tidak menjamin dapat menaikan koefisien relibilitas. Hal ini juga sejalan dengan pendapat Croker dan Algina bahwa meskipun *reward score* dapat mengestimasi kemampuan dengan baik, tidak mendukung anggapan bahwa dapat meningkatkan reliabilitasnya.

Oleh karenanya jika diskor dengan teknik *reward*, semakin banyak jawaban yang dikosongkan, maka akan diperoleh relibilitas yang kecil. Sebaliknya semakin banyak melakukan tebakan, maka reliabilitas akan cenderung lebih tinggi atau meningkat. Dengan demikian perolehan skor yang menunjukkan keterpercayaan skor siswa yang sesuai dengan kemampuan yang dimilikinya ditunjukan dengan koefisien reliabilitas yang cenderung rendah.

Selain itu, jika dianalisis dari hasil, dengan teknik *reward* pengosongan jawaban itu tidak hanya dilakukan oleh siswa yang memiliki kemampuan rendah saja. Namun semua level kemampuan memilih lebih mengosongkan jawaban terhadap soal yang belum pasti kebenaran jawabannya (ragu-ragu)

2. Ada perbedaan reliabilitas teknik penskoran *Number-right score* dan *punishment score*

Teknik penskoran *punishment score* adalah penskoran dengan cara menjumlahkan jawaban benar dan jawaban salah oleh setiap siswa pada masing-masing butir soal. Butir jawaban benar diberi skor 1, jawaban salah diberi pengurangan skor sebesar -0,25, dan jawaban dikosongkan diberi skor 0. Jadi maksimum skor yang diperoleh siswa adalah 25. Pengurangan jawaban salah merupakan hukuman terhadap tebakan. Sedangkan *number-right* merupakan cara menjumlahkan jawaban benar saja. Jawaban benar diberi skor 1, jawaban salah dan dikosongkan diberi skor 0. Dalam penelitian ini, siswa yang dikenakan teknik penskoran *punishment score* sekaligus *number-right* berjumlah 163 siswa.

Tabel 3. **Hasil Uji Hipotesis 2**

Paired Samples Test

	Paired Differences					t	Df	Sig. (2-tailed)
	Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference				
				Lower	Upper			
Pair 1 NR - P	,0320	,022016	,004019	,023779	,040221	7,961	29	,000

Hasil analisis diperoleh t=7,961, sig=0,00. Dengan menggunakan taraf signifikansi 5% maka $p\ value < 0{,}05$, dengan demikian disimpulkan bahwa H_0 ditolak. Ini mengindikasikan bahwa ada perbedaan reliabilitas antara teknik penskoran *number-right score* dan *punishment score*. Jika dilihat dari nilai rata-rata reliabilitas, terlihat bahwa *mean* reliabilitas *number right score* lebih besar dari *mean* reliabilitas *punishment score*, yaitu 0,671 > 0,639

Sebelumnya dibahas, dengan teknik penskoran *number right score*, bisa saja soal sukar direspon benar oleh siswa yang tidak memiliki kecukupan pengetahuan, dan dijawab salah oleh yang memiliki kecukupan pengetahuan. Jawaban benar yang diberikan oleh peserta tes yang memiliki kemampuan rendah dan jawaban salah oleh kemampuan tinggi tentu berakibat pada reliabilitas yang dihasilkan, yaitu akan diperoleh reliabilitas yang rendah. Karena jika peserta tes yang tidak memiliki kecukupan pengetahuan dapat menjawab butir soal yang sukar dengan benar, maka dapat diindikasikan peserta tersebut melakukan penebakan. Sehingga perlu alternatif teknik lain untuk mengurangi penebakan itu. Selain *reward*, teknik *Corection for geussing* adalah *punishment score*. Teknik ini bertujuan untuk meminimalisir *guessing* agar reliabilitas lebih baik. Tapi analisis menunjukan bahwa rata-rata reliabilitas dengan teknik *number right score* lebih tinggi daripada *punishment score*. Hasil ini juga sejalan dengan penelitian yang dilakukan oleh Venny (2009) pada penelitiannya yang berjudul perbandingan daya beda dan reliabilitas tes pilihan ganda berdasarkan model penskoran siswa menyimpulkan bahwa model penskoran komposit memiliki reliabilitas lebih besar dari model penelitian pinalti.

Tapi pada penelitian ini, bukan berarti teknik *number right score* lebih baik daripada *punishment score*. Perlu diingat, bahwa dalam penelitian ini perolehan skor pada *number right score* merupakan hasil respon dari instrumen dengan *punishment* dengan memberi skor 1 pada jawaban benar dan skor 0 pada respon lainnya, sehingga skor yang dihasilkan pada teknik *number right score* benar-benar skor yang sesuai

dengan kemampuan pengikut tes. Artinya pengikut tes yang memiliki kecukupan kemampuan akan merespon tes dengan benar, sebaliknya pengikut tes yang tidak memiliki kecukupan pengetahuan akan merespon tes dengan salah

Selain itu, pada teknik penskoran *punishment score*, jika pengikut tes tidak melakukan tebakan (jawaban dikosongkan) diberi skor 0. Bila dianalogikan reliabilitas yang dihasilkan oleh *number-right* skor sebagai reliabilitas *punishment score,* dengan skor 0 merupakan jawaban yang tidak melakukan tebakan, maka tidak melakukan tebakan pada *punishment score* akan meningkatkan reliabilitas.

Dengan demikian pada teknik penskoran *punishment score*, jika siswa tidak melakukan tebakan maka dapat menaikan koefisien relibilitas. Hal ini juga sejalan dengan pendapat Croker dan Algina bahwa dengan mengoreksi *guessing* dapat meningkatkan reliabilitas dan validitasnya.

Oleh karenanya jika diskor dengan teknik *punishment*, semakin banyak jawaban yang dikosongkan, maka akan diperoleh relibilitas yang cenderung baik. Sebaliknya semakin banyak melakukan tebakan, maka reliabilitas akan cenderung kecil. Dengan demikian perolehan skor yang menunjukkan keterpercayaan skor siswa yang sesuai dengan kemampuan yang dimilikinya ditunjukan dengan koefisien reliabilitas yang cenderung tinggi.

Selain itu, jika dianalis dari hasil, dengan teknik *punishment score* kehati-hatian tidak berlaku pada semua level kemampuan. Pada level kemampuan sedang dan rendah tetap saja melakukan penebakan. Kehati-hatian hanya belaku pada siswa yang memiliki kemampuan tinggi saja. Banyak faktor yang mempengaruhinya, salah satunya adalah dengan adanya remedial. Karena siswa beranggapan kalaupun toh nilainya rendah nanti juga bisa diperbaiki kembali.

3. Pada kelompok teknik penskoran *reward score*, tidak ada perbedaan reliabilitas antara ukuran sampel 30 dan 40

Ukuran sampel 30 dan 40 maksudnya adalah banyaknya orang yang digunakan untuk menghitung setiap koefisien reliabilitas.

Tabel 4. **Hasil Uji Hipotesis 3**
Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference	
									Lower	Upper
SKOR	Equal variances assumed	1,204	,277	-1,165	58	,249	-,018433	,015825	-,050111	,013244
	Equal variances not assumed			-1,165	56,550	,249	-,018433	,015825	-,050128	,013261

Dari hasil analisis diperoleh t=-1,165, sig=0,249. Dengan menggunakan taraf signifikansi 5% maka $p\ value > 0{,}05$, dengan demikian H_0 diterima. Dengan menerima H_0, disimpulkan bahwa pada kelompok *reward*, tidak ada perbedaan antara ukuran sampel 30 dan 40. Dengan teknik ini, jawaban salah akan diberi *score* nol, jawaban benar diberi skor 1, sedangkan jawaban yang dikosongkan diberi skor 0,2. Jawaban salah diasumsikan bahwa siswa melakukan penebakan. Sedangkan jawaban yang dikosongkan diasumsikan bahwa siswa benar-benar tidak bisa menjawab soal tersebut atau siswa malas mencarinya karena dengan mengosongkanpun akan mendapat penambahan *score.* Kedua hal tersebut berakibat pada reliabilitas yaitu keterpercayaan hasil siswa diragukan. Formula relibilitas *Cronbach Alpha* adalah $\rho_\alpha = \frac{N}{N-1}\left(1 - \frac{\sum \sigma_i^2}{\sigma_A^2}\right)$, dimana $\sum \sigma_i^2$ = jumlah seluruh variansi butir, dan $\sum \sigma_A^2$ = variansi skor responden. Pada teknik penskoran *reward score* jika menebak, maka pada setiap butir tersebut diberi *score* nol. Sedangkan dikosongkan diberi *score* 0,2. Pemberian *score* yang demikian berakibat pada jumlah variansi butir dan variansi totalnya. Namun variansi itu berbanding setara, artinya jika $\frac{\sum \sigma_i^2}{\sigma_A^2}$ maka angka yang keluar akan kecil karena pembilang akan selalu lebih kecil dari penyebutnya. Kemudian jika hasil pembagian tersebut dimasukan pada formula $1 - \frac{\sum \sigma_i^2}{\sigma_A^2}$, maka angka yang keluar cenderung besar. Oleh karenanya reliabilitas akan besar pula. Apalagi *instruction* pada teknik ini diberikan dengan tujuan agar siswa memiliki keputusan untuk memilih mengosongkan atau menebak jawaban soal. Keputusan tersebut berkaitan dengan faktor psikologis yang berakibat pada sikap pengikut tes terhadap pengerjaannya sehingga terjadi perbedaan keputusan karena *Trait* masing-masing individu

berbeda. Selain itu, hal tersebut juga berkaitan dengan variasi sebaran *score* siswa, artinya jika siswa melakukan contekan maka respon akan cenderuang sama satu dengan yang lainya. Pada uji ini tidak terdapat perbedaan reliabilitas dengan menggunakan ukuran responden 30 dan 40. Ini mengindikasikan bahwa tidak ada efek dari penambahan ukuran sampel, yaitu sebaran variasi data tidak beragam sehingga reliabilitas juga ekuivalen.

4. Pada kelompok teknik penskoran *punishment score*, tidak ada perbedaan reliabilitas antara ukuran sampel 30 dan 40

Ukuran sampel 30 dan 40 maksudnya adalah banyaknya orang yang digunakan untuk menghitung setiap koefisien reliabilitas.

Tabel 5. Hasil uji hipotesis 4

Independent Samples Test										
		Levene's Test for Equality of Variances		t-test for Equality of Means						
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference	
									Lower	Upper
SKOR	Equal variances assumed	,338	,563	1,427	58	,159	,024933	,017471	-,010038	,059905
	Equal variances not assumed			1,427	57,035	,159	,024933	,017471	-,010050	,059917

Dari hasil analisis diperoleh t=1,427, sig= 1,59. Dengan menggunakan taraf signifikansi 5% maka $p\ value > 0{,}05$, dengan demikian H_0 diterima. Dengan menerima H_0, disimpulkan bahwa pada kelompok *punishment*, tidak ada perbedaan antara ukuran sampel 30 dan 40

Koefisien reliabilitas konsistensi internal *Cronbach Alpha* ditentukan oleh kovariansi diantara sekor butir. Jika semua butir konsisten setara maka kovariansi menjadi besar dan koefisien reliabilitas menjadi tinggi. Formula relibilitas *Cronbach Alpha* adalah $\rho_\alpha = \frac{N}{N-1}\left(1 - \frac{\sum \sigma_i^2}{\sigma_A^2}\right)$, dimana $\sum \sigma_i^2$ = jumlah seluruh variansi butir, dan $\sum \sigma_A^2$ = variansi skor responden. Pada teknik penskoran *punishment score* semakin banyak menebak, maka pada setiap butir tersebut mengalami pengurangan, pengurangan tersebut berakibat pada semakin kecilnya

skor total yang diperoleh. Hal ini berakibat pada jumlah variansi butir semakin besar sedangkan variansi totalnya semakin kecil. Dengan demikian, jika $\frac{\sum \sigma_i^2}{\sigma_A^2}$ maka angka yang keluar akan besar karena pembilang akan selalu lebih besar dari penyebutnya. Kemudian jika hasil pembagian tersebut dimasukan pada formula $1 - \frac{\sum \sigma_i^2}{\sigma_A^2}$, maka angka yang keluar cenderung kecil. Maka reliabilitasnya akan kecil pula. Jika ukuran responden ditambah dalam penghitungan relibilitas, maka hasilnya tidak akan jauh berbeda dengan ukuran responden yang digunakan sebelumnya. Apalagi *instruction* pada teknik ini diberikan dengan tujuan agar siswa memiliki keputusan untuk memilih mengosongkan atau menebak jawaban soal. Keputusan tersebut berkaitan dengan faktor psikologis yang berakibat pada sikap pengikut tes terhadap pengerjaannya sehingga terjadi perbedaan keputusan karena *Trait* masing-masing individu berbeda. Hasil uji menunjukan bahwa tidak ada perbedaan, hal ini kemungkinan disebabkan pada teknik peskoran *punishment score* sebaran respon siswa cenderung tidak bervariasi. Banyak hal yang mengindikasikan kehomogenan data yang diperoleh. Dimungkinkan siswa untuk saling contek, belum terbiasa dengan teknik *punishment* ataupun materi yang belum dikuasai pada kelompok tersebut sehingga terjadi kesalahan yang sama.

5. Kesimpulan

Kesimpulan hasil penelitian ini adalah:

1. Ada perbedaan yang *significant* reliabilitas tes pilihan ganda hasil belajar Matematika antara teknik penskoran *number-right score* dengan *reward score*
2. Ada perbedaan yang *significant* reliabilitas tes pilihan ganda hasil belajar Matematika antara teknik penskoran *number-right score* dengan *punishment score*
3. Tidak ada perbedaan reliabilitas tes pilihan ganda hasil belajar Matematika kelompok *reward* dengan ukuran sampel 30 dan 40
4. Tidak ada perbedaan reliabilitas tes pilihan ganda hasil belajar Matematika kelompok *punishment* dengan ukuran sampel 30 dan 40

5. Daftar Pustaka

Ali, Mohammad. 2011. *Melakukan Riset Prilaku dan Sosial.* Pustaka Cendikia Utama: Bandung

Algina dan Crocker. 1986. *Introduction to Clasisical And Modern Test Theory.*

Anastasi, Anne. 2007 . *Tes Psikologi*. PT. Indeks : Jakarta

Azwar, Saifuddin. 2012a. *Reliabilitas dan Validitas*. Pustaka Pelajar : Yogyakarta

_____ 2012b. *Dasar-dasar Pikometri*. Pustaka Pelajar :Yogyakarta.

_____ 2011. *Tes Prestasi*. Pustaka Pelajar :Yogyakarta.

Budiyono. 2000. *Statistik Dasar untuk Penelitian.* FKIP MIPA : UNS

Budi B, Yoga. 2013. Pengaruh *jumlah Alternatif Jawaban dan Teknik penskoran terhadap reliabilitas ter IPA terpad.* TESIS UNJ :Tidak diterbitkan.

Busnawir. 2006. *Pengaruh Model Penskoran Terhadap Kesetabilan Reliabilitas Hasil Pengukuran Skala Sikap dengan Mempertimbangkan Varians Usia.* Disertasi UNJ :Tidak Diterbitkan

Dalyono, M. 2010. *Psikologi Pendidikan*. Rineka Cipta: Jakarta

Furqon. 2009. *Statistika Terapan Untuk Penelitian*. Alfa Beta : Bandung

Jazuli. Akhmad. 2009. *Prosiding Matematika*. Sumber :eprints.uny.ac.id/7025/1/P11-Akhmad%20Jazuli.pdf. didownload 1 Maret 2014, Pukul 20.00 WIB

Kaplan, M Robert dan Saccuzzo, Dennis P. 2012. *Pengukuran Psikologi : Prinsip,penerapan, dan Isu*. Salemba Humanika: Jakarta

Kementerian Pendidikan dan Kebudayaan.2013. *Matematika kelas X Kurikulum 2013 : Buku Guru*. *E-Book* Kementerian Pendidikan dan Kebudayaan: Jakarta

Krathwol, David R dan Anderson, Lorin W (editor). 2010. *Kerangka Landasan untuk pengajaran Pembelajaran dan assesment*. Pustaka Pelajar: Yogyakarta

Lord, Frederic M dan Novick, Melvin R. 1968. *Statistical Theories of Mental Test Score*. Massachusett. Addison-Wesley Publishing company, Inc.

Marwanta, dkk (anggota Ikapi). 2008). *Mathematics For Senior High School Year X*. Yudhistira

Naga, Dali S. 2013. *Teori Sekor Dalam Pengukuran Mental*. PT Nagarani Citrayasa: Jakarta

Naga, Dali S. 2008. *Probabilitas dan Sekor pada Hipotesis Statistika*. UPT Penerbitan Universitas Tarumanegara

Rasyid, Harun, dkk. 2007. *Penilaian Hasil Belajar*. CV Wacana Prima :Bandung

Riska I, Venny. 2009. *Perbandingan Daya Beda dan Reliabilitas Tes Pilihan Ganda Berdasarkan Model Pensekoran*. Jakarta :UNJ

Saefudin, Abdul A dan Kusumaningrum, Maya. 2012. *Mengoptimalkan Kemampuan Berfikir Matematika Malalui Pemecahan Masalah Matematika*. Prosiding

Santrock, John W. 2003. *Andolescence Perkembangan Remaja*. Erlangga : Jakarta

Sigit Suprijanto, dkk (anggota Ikapi). 2009. *Mathematics For Senior High School Year XI*. Yudhistira

Suharsaputra, Uhar. 2012. *Metode Penelitian : Kuantitatif, Kualitatif, dan Tindakan*. PT Refika Aditama: Bandung

Sukardi, Dewa Ketut dan Kusmawati, Nila. 2009. *Analisis Tes Psikologi Teori dan Prakti*k. Rhineka Cipta : Jakarta

Surapranata, Sumarna. 2009. *Analisis, Validitas, Reliabilitas, dan Interpretasi Hasil*. PT Remaja Rosdakarya: Bandung.

Suprijono, Agus. 2009. *Cooperative Learning: Teori dan Aplikasi Paikem*. Pustaka Belajar: Yogyakarta

Susetyo, Budi. 2011. *Menyusun Tes Hasil Belajar*. CV Cakra: Bandung

Suryabrata, Sumadi. 2002. *Pengembangan Alat Ukur Psikologis*. Andi: Yogyakarta.

Throndike et al, Robert M. 1991. *Measurement and Evaluation In Psychology and Education*. New York Macmilan Publishing Company

Uyanto, Stanislaus S.2009. *Pedoman Analisi data dengan SPSS*.Graha Ilmu: Yogyakarta

Venny Riska, Indriyani. (2009). *Perbandingan Daya Beda Dan Reliabilitas Tes Pilihan Ganda Berdasarkan Model Pensekoran.* Jakarta: UNJ

Widiatmoko. 2009. *Pengaruh Metode Penyekoran Pada Stabilitas Koefisisen Reliabilitas Sekor Tes Objektif Pilihan Ganda Ditinjau Dari Keragaman Intelegensi Peserta Tes*. Disertasi UNJ : Tidak Diterbitkan.

Yuri R, Levita. 2011. *Pengaruh Metode Pensekoran dan Risk Talking Level Terhdap Relibilitas Tes Matematika*. Disertasi UNJ : Tidak diterbitkan

Zainul, Asmawi. 1993. *Penilaian Hasil Belajar*. PAU-PPAI Universitas Terbuka: Jakarta

http://digilib.uinsuka.ac.id/8036/1/EVA%20LATIPAH%20STRATEGIPENGENALAN%20POTENSI%20ANAK.pdf. Didownload 5 Februari 2014, pukul 09.00 WIB