

PENGEMBANGAN INSTRUMEN TES UNTUK MENGUKUR KEMAMPUAN PENALARAN STATISTIK MAHASISWA TADRIS MATEMATIKA

Budi Manfaat, Siti Nurhairiyah

Jurusan Pendidikan Matematika, Fakultas Tarbiyah, IAIN Syekh Nurjati Cirebon,
Jalan Perjuangan By Pass Cirebon 45132, Indonesia
Telepon : +62 231 481264

ABSTRAK

Evaluasi pembelajaran tidak lepas dari penggunaan instrumen. Instrumen ini mempunyai peranan yang penting dalam mengetahui sejauh mana kemampuan peserta didik dalam memahami materi. Instrumen yang sering diberikan dosen pada pembelajaran statistika berbentuk tes essay. Instrumen ini sesuai untuk mengukur kemampuan penalaran statistik, akan tetapi memiliki unsur subjektivitas yang tinggi. Sementara untuk bentuk pilihan ganda memiliki unsur objektivitas yang tinggi, tetapi unsur menebak jawaban peserta didik juga tinggi dan kurang dapat mengukur kemampuan peserta. Oleh karena itu perlu dikembangkan bentuk tes pilihan ganda yang memiliki kualitas baik dan dapat mengukur kemampuan penalaran statistik mahasiswa.

Penelitian ini bertujuan: (1) Untuk menemukan instrumen yang dapat mengukur kemampuan penalaran statistik mahasiswa ditinjau dari karakteristik instrumen yang baik. (2) Untuk mengetahui hasil analisis kemampuan penalaran statistik dengan menggunakan instrumen yang telah dikembangkan. Instrumen tes untuk mengukur kemampuan penalaran statistik adalah instrumen tes yang meliputi aspek penalaran terhadap ukuran pemusatan, penalaran terhadap ketidakpastian, penalaran terhadap sampel, dan penalaran terhadap perkumpulan. Instrumen tes tersebut berupa pilihan ganda yang memiliki tingkatan soal analisis (C4), sintesis (C5), dan evaluasi (C6).

Penelitian ini menggunakan metode penelitian dan pengembangan (*Research & Development*). Adapun tahapannya meliputi studi pendahuluan dan pengembangan model. Pada tahap pengembangan model melalui tahap validasi isi, uji coba terbatas (dengan 30 mahasiswa dari semester II, IV dan VI), dan uji coba luas (dengan 171 mahasiswa dari semester II, IV dan VI). Semua mahasiswa dipilih secara random.

Hasil penelitian menunjukkan bahwa: (1) Instrumen yang dikembangkan memiliki karakteristik yang baik dari segi validitas dengan indeks validitas isi (CVI) sebesar 0,94 dan koefisien reliabilitas sebesar 0,68. Adapun dari segi daya pembeda, butir yang berkategori sangat baik sebanyak 11 butir dan yang baik sebanyak 3 butir, dari segi tingkat kesukaran semua butir berkategori sedang, dari segi korelasi skor butir dengan skor total yang signifikan sebanyak 12 butir, dan dari kualitas distraktor yang efektif sebanyak 13 butir. (2) Hasil analisis kemampuan penalaran statistik dengan menggunakan instrumen yang telah dikembangkan menunjukkan bahwa 99% peserta tes memiliki nilai estimasi kemampuan/*ability* yang berkategori cukup yaitu antara -1,5 sampai dengan kurang dari 1,5, sementara hanya 1% peserta tes yang memiliki nilai estimasi kemampuan/*ability* dengan kategori baik yaitu antara 1,5 sampai dengan kurang dari 3.

Kata Kunci : instrumen tes, research & development

PENDAHULUAN

Dalam setiap proses pembelajaran di jenjang manapun akan selalu ada proses evaluasi atau penilaian. Proses evaluasi tersebut merupakan hal yang sangat penting bagi perkembangan kualitas pendidikan di semua negara khususnya Indonesia. Hal ini dijelaskan dalam UU SISDIKNAS Nomor 20 Tahun 2003 Bab XVI tentang Evaluasi, Akreditasi dan Sertifikasi pada Pasal 57 Ayat 1 “Evaluasi dilakukan dalam rangka pengendalian mutu pendidikan secara nasional sebagai bentuk akuntabilitas penyelenggara pendidikan kepada pihak – pihak yang berkepentingan.” Selain itu, tujuan diadakannya evaluasi menurut Sudijono (2008:16) adalah :

“Untuk memperoleh data pembuktian yang akan menjadi petunjuk sampai dimana hasil tingkat kemampuan dan tingkat keberhasilan peserta didik dalam pencapaian tujuan-tujuan pembelajaran setelah mereka menempuh proses pembelajaran.”

Dalam proses evaluasi, seorang pendidik tak lepas dari yang namanya penggunaan instrumen. Menurut Arifin (2012:68) instrumen itu mempunyai fungsi dan peran yang sangat penting dalam rangka mengetahui keefektifan proses pembelajaran. Selain itu, Mulyadi (2010:2) mengemukakan bahwa proses evaluasi mencakup dua hal yakni pengukuran dan tes. Ketika melakukan evaluasi, maka pendidik harus melakukan pengukuran yang di dalamnya juga harus menggunakan alat yang biasa disebut tes. S.Hamid Hasan dalam Arifin (2012:3) menerangkan bahwa tes adalah alat pengumpulan data yang dirancang secara khusus.

Statistik merupakan salah satu mata kuliah yang harus diberikan pendidik kepada peserta didik di jenjang mahasiswa terutama di Jurusan Tadris Matematika. Salah satu tujuan pembelajaran statistik adalah membentuk dan mengembangkan kemampuan penalaran statistik. Menurut Garfield (2003) kemampuan ini memperkenalkan seseorang terhadap konsep-konsep dasar dan logika penalaran statistik dan memberikan pengantar tingkat kemampuan praktis untuk memilih, menghasilkan, dan benar menafsirkan metode deskriptif dan inferensial yang tepat. Materi statistik yang diajarkan dirancang untuk membantu mahasiswa memahami dan menangani ketidakpastian, variabilitas dan informasi statistik yang didapat. Selain itu, tentu saja membantu mahasiswa memperoleh penghargaan untuk aplikasi beragam statistik dan relevansinya dengan kehidupan mereka dan bidang studi.

Manfaat dari kemampuan penalaran statistik dapat diterapkan tidak hanya dalam pendidikan melainkan dalam segala bidang seperti dalam bidang ekonomi dan bisnis, dalam bidang kependudukan, ilmu administrasi, dan sebagainya. Sebagai contoh, keberhasilan Jepang dalam menerapkan ilmu statistika terutama ilmu peluang (probabilitas) dalam mendesain dan memasarkan produk-produknya seperti motor, barang elektronik dan sebagainya. Menurut Boediono dan Kaster (2004) prestasi itu dicapai karena keberhasilan pendidikan di Jepang dalam mata pelajaran statistika yang lebih menekankan dalam hal penalaran. Materi ini diberikan secara luas sejak masyarakat Jepang menginjak pendidikan sekolah menengah atas sampai perguruan tinggi.

Begitu pentingnya kemampuan penalaran statistik bagi mahasiswa sehingga dosen harus berupaya mengembangkan kemampuan penalaran statistik mahasiswanya. Salah satu cara untuk mengembangkan kemampuan penalaran statistik adalah dengan melatih mahasiswa mengerjakan instrumen berupa tes-tes yang berhubungan dengan kemampuan penalaran statistik. Dengan begitu mahasiswa akan lebih mengembangkan kemampuan penalaran statistik. Sehingga kualitas mahasiswa meningkat, bahkan jika lebih banyak mahasiswa yang mengembangkan kemampuan penalaran statistiknya maka kualitas pendidikannya pun meningkat. Dengan mengoptimalkan kemampuan penalaran artinya mensyukuri anugerah terbesar dari Allah SWT yakni akal dan pikiran. Karena kedua hal inilah yang membedakan antara manusia dengan makhluk lainnya. Hal ini sesuai dengan firman Allah SWT dalam Surat Az-Zumar ayat 9 yang artinya:

“.....Katakanlah apakah sama orang-orang yang mengetahui dengan orang-orang yang tidak mengetahui? Sesungguhnya orang yang berakal lah yang dapat menerima pelajaran”.

Berdasarkan hasil survei lapangan di jurusan Tadris Matematika IAIN Syekh Nurjati Cirebon, kemampuan penalaran statistik mahasiswa bervariasi. Sebagian mahasiswa masih mengalami kesulitan ketika diberikan soal penalaran. Selain itu, mahasiswa kurang menerapkan pengetahuan yang telah didapat dari pendidik dalam kehidupan sehari-hari sedangkan pengetahuan tersebut sangat penting untuk diterapkan. Ketika mahasiswa mempelajari materi populasi dan sampel kemudian dihadapkan pada suatu permasalahan tertentu, maka langkah yang mesti ditempuh adalah dengan mengambil keputusan berdasarkan pertimbangan sampel yang lebih luas dan langkah tersebut merupakan salah satu bagian dari proses bernalar dalam mengidentifikasi sebab terjadinya suatu masalah sedang dilakukan. Akan tetapi sebagian mahasiswa belum melakukannya dan tidak menerapkan pengetahuan statistik yang telah mereka dapatkan dalam mengambil keputusan tersebut. Informasi ini diperoleh melalui kegiatan diskusi kelompok kecil yang diadakan ketika survei lapangan.

Pada saat evaluasi atau pengujian, pendidik kurang memberikan instrumen yang menuntut kemampuan penalaran statistik. Instrumen yang biasa diberikan kepada mahasiswa pada mata kuliah statistik yakni berupa tes yang berbentuk uraian. Tes tersebut seringkali terfokus pada akurasi perhitungan data statistik, aplikasi yang benar dari formula atau kebenaran dari grafik dan diagram. Selain itu, penilaian yang dilakukan pendidik dipengaruhi subjektivitas pendidik itu sendiri. Sehingga hasil evaluasi mahasiswa baik yaitu dengan nilai yang tinggi, akan tetapi tidak sesuai dengan pola pikir mahasiswa. Begitu mudahnya mahasiswa mendapat nilai yang baik atau tinggi dengan tanpa menambah kualitas yang ada pada diri mahasiswa atau mahasiswa tanpa berpikir lebih dalam.

Menurut delMas (2002) bentuk tes essay atau uraian itu sesuai untuk mengukur kemampuan penalaran statistik, karena harus berlandaskan “mengapa (*why*) dan bagaimana (*how*)”. Akan tetapi, bentuk tes essay itu memiliki kelemahan seperti yang terjadi di lapangan. Kemudian bagaimanakah caranya jika yang digunakan untuk meningkatkan kemampuan penalaran statistik mahasiswa adalah bentuk tes pilihan ganda?

Bentuk tes pilihan ganda memiliki kelebihan yaitu dapat memberikan hasil yang objektif yang tinggi dan dapat dipercaya, dalam pengoreksian jawabannya tidak sulit, dan lebih representatif dalam hal mencakup dan mewakili materi yang telah diajarkan (Sudijono, 2008:133). Akan tetapi, bentuk pilihan ganda juga memiliki kelemahan yaitu kurang dapat mengukur atau mengungkap proses berpikir yang tinggi atau mendalam, terbuka kemungkinan bagi *testee* untuk bermain spekulasi dan menebak jawaban, dan pembuatan butir soalnya tidak semudah tes essay (Sudijono, 2008: 134).

Untuk mengatasi kelemahan yang dimiliki tes pilihan ganda, maka perlu dikembangkan agar bentuk pilihan ganda itu tetap memiliki kualitas yang baik. Menurut Azwar (2010:74), bentuk pilihan ganda juga dapat mengungkap kemampuan berpikir tingkat tinggi, karena bentuk ini memiliki variasi tingkat penguasaan yang paling luas, mulai dari yang sederhana sampai kepada yang paling tinggi. Selain itu, bentuk tes pilihan ganda juga dapat digunakan untuk mengukur kemampuan berpikir tinggi seperti kemampuan penalaran statistik dengan syarat butir-butir soalnya harus memiliki tingkatan soal antara C4 sampai C6 yaitu tingkatan analisis, sintesis dan evaluasi. Hal ini seperti yang diungkapkan dalam Taxonomy Bloom dalam Betsy Moore & Tood Stanley (2010:35) bahwa “*analysis question, sintesis, question, and evaluationquestion can also be asked using a multiple choice format*”. Artinya soal yang berbentuk analisis, sintesis dan evaluasi dapat digunakan dengan bentuk soal pilihan ganda. Dengan demikian tes untuk mengukur kemampuan penalaran statistik dapat direpresentasikan dalam bentuk tes pilihan ganda.

Berdasarkan uraian di atas, maka diperlukan instrumen tes yang dapat mengukur kemampuan penalaran statistik mahasiswa, sehingga memotivasi peneliti untuk melakukan penelitian tentang “Pengembangan Instrumen Tes untuk Mengukur Kemampuan Penalaran Statistik Mahasiswa Tadris Matematika IAIN Syekh Nurjati Cirebon”.

METODE DAN SUBJEK PENELITIAN

A. Metode Penelitian

Metode yang digunakan dalam penelitian ini adalah metode R & D (*Research and Development*) atau metode penelitian dan pengembangan. Secara sederhana penelitian dan pengembangan (*Research and Development*) menurut Putra (2011:67) dapat didefinisikan sebagai metode penelitian yang secara sengaja, sistematis, bertujuan/diarahkan untuk mencari temuan, merumuskan, memperbaiki, mengembangkan, menghasilkan, menguji keefektifan produk, model, metode/strategi/cara, jasa, prosedur tertentu yang lebih unggul, baru, efektif, efisien, produktif dan bermakna.

Secara garis besar langkah penelitian dan pengembangan yang dimodifikasi oleh Sukmadinata (2012:184) berasal dari sepuluh langkah penelitian dan pengembangan yang dimodifikasi dari sepuluh langkah penelitian dan pengembangan dari Borg dan Gall. Langkah penelitian yang dimodifikasi tersebut terdiri dari tiga tahap yaitu :

1. Studi pendahuluan yang meliputi studi kepustakaan, survei lapangan dan penyusunan produk awal.
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.

Metode penelitian yang akan dilakukan dalam penelitian ini mengacu kepada pendapat Sukmadinata. Akan tetapi penelitian yang dilakukan hanya sampai pada tahap pengembangan model seperti yang dijelaskan Sukmadinata (2012:187) bahwa untuk peneliti dari program S2 atau penyusunan tesis, kegiatan penelitian dan pengembangan dapat dihentikan sampai dihasilkan draf final tanpa pengujian hasil yakni sampai tahapan pengembangan model.

B. Subjek Penelitian

Subjek penelitian dalam tahap pengembangan instrumen tes ini adalah mahasiswa jurusan Tadris Matematika IAIN Syekh Nurjati Cirebon yang berada pada semester II, IV dan VI yang mengikuti kegiatan pembelajaran pada tahun 2013. Mahasiswa tersebut dibagi menjadi dua kelompok uji coba yakni kelompok ujicoba terbatas sebanyak 30 mahasiswa dan kelompok uji coba luas sebanyak 171 mahasiswa. Semua mahasiswa tersebut dipilih secara acak (*random*).

HASIL PENELITIAN DAN ANALISIS DATA

A. Studi Pendahuluan

Berdasarkan hasil studi pendahuluan yang telah dilakukan, diperoleh informasi mengenai konsep-konsep atau teori-teori yang berkenaan dengan produk yang akan dikembangkan seperti pengertian dan kriteria suatu tes, langkah-langkah-langkah pengembangan instrumen, dan mengenai penalaran statistik. Selain itu diperoleh juga informasi yang paling penting yaitu jenis tes yang biasa digunakan oleh pendidik serta indikator pembelajaran.

Jenis tes yang biasa digunakan adalah bentuk essay atau uraian. Tes tersebut biasanya berbentuk pengaplikasian dari suatu formula. Tes yang berbentuk essay atau uraian sebenarnya sesuai untuk mengukur kemampuan penalaran statistik dikarenakan bentuk tersebut dapat mengungkap kemampuan *testee* dalam memahami berbagai macam konsep. Selain itu, menurut delMas (2002) bentuk tes untuk mengukur kemampuan penalaran statistik harus berlandaskan “mengapa (*why*) dan bagaimana (*how*)”, sehingga bentuk tes essay sesuai. Akan tetapi, bentuk tes essay itu memiliki kelemahan diantaranya terdapat kecenderungan *tester* yang lebih banyak bersifat subjektif, kurang menampung atau mencakup luasnya materi yang diberikan kepada *testee*, dan membutuhkan waktu yang lama dalam pengkoreksian jawabannya (Sudijono, 2008:103).

Dalam penelitian ini, bentuk tes yang akan dikembangkan adalah tes objektif yang berbentuk pilihan ganda. Bentuk tes ini memiliki kelebihan yaitu dapat memberikan hasil yang objektif yang tinggi dan dapat dipercaya, dalam pengoreksian jawabannya tidak sulit, dan lebih representatif dalam hal mencakup dan mewakili materi yang telah diajarkan (Sudijono, 2008:133). Akan tetapi, bentuk pilihan ganda juga memiliki kelemahan yaitu kurang dapat mengukur atau mengungkap proses berpikir yang tinggi atau mendalam, terbuka kemungkinan

bagi *testee* untuk bermain spekulasi dan menebak jawaban, dan pembuatan butir soalnya tidak semudah tes essay (Sudijono, 2008: 134).

Untuk mengatasi kelemahan yang dimiliki tes pilihan ganda, maka perlu dikembangkan agar bentuk pilihan ganda itu tetap memiliki kualitas yang baik. Menurut Azwar (2010:74), bentuk pilihan ganda juga dapat mengungkap kemampuan berpikir tingkat tinggi, karena bentuk ini memiliki variasi tingkat penguasaan yang paling luas, mulai dari yang sederhana sampai kepada yang paling tinggi. Selain itu, bentuk tes pilihan ganda juga dapat digunakan untuk mengukur kemampuan berpikir tinggi seperti kemampuan penalaran statistik dengan syarat butir-butir soalnya harus memiliki tingkatan soal antara C4 sampai C6 yaitu tingkatan analisis, sintesis dan evaluasi. Hal ini seperti yang diungkapkan dalam Taxonomy Bloom dalam Betsy Moore & Tood Stanley (2010:35) bahwa “*analysis question, sintesis, question, and evaluationquestion can also be asked using a multiple choice format*”. Artinya soal yang berbentuk analisis, sintesis dan evaluasi dapat digunakan dengan bentuk soal pilihan ganda.

Dalam Taksonomi Bloom, pada tingkatan soal C4 (analisis) kata kerja operasional yang digunakan adalah menganalisis, mengaudit, memecahkan, menyimpulkan, menelaah, mengaitkan, dan mendeteksi. Untuk tingkatan soal C5 (sintesis) kata kerja operasional yang digunakan adalah menggabungkan, merumuskan, mengkategorikan, menciptakan, merancang, dan menyusun. Sementara itu, untuk tingkatan C6 (evaluasi) kata kerja operasional yang digunakan adalah menimbang, mengkritik, membandingkan, menyimpulkan, menilai, memprediksi, dan menafsirkan (Azwar, 2010:64). Dengan demikian tes untuk mengukur kemampuan penalaran statistik dapat direpresentasikan dalam bentuk tes pilihan ganda. Langkah selanjutnya yaitu penyusunan produk dengan cara membuat kisi-kisi instrumen (dapat dilihat pada lampiran A.1) dan menyusun instrumen tes sebanyak 20 butir soal (dapat dilihat pada lampiran A.2).

B. Pengembangan Model

Pengembangan model yang dimaksud adalah pengembangan instrumen tes dalam penelitian ini. Pada tahap pengembangan instrument tes ini diperoleh informasi mengenai kualitas instrumen tes yang dikembangkan. Tahap pengembangan model meliputi tahap estimasi validitas dan tahap uji coba empirik. Tahap validasi instrumen dilakukan untuk mengetahui bahwa instrumen yang dibuat telah sesuai dengan indikator yang ada. Tahapan ini diperoleh dari hasil analisis secara teoritik yaitu dengan menggunakan pertimbangan tim ahli. Selanjutnya tahap uji coba empirik meliputi tahap uji coba terbatas dan tahap uji coba luas. Melalui tahapan ini dapat diketahui estimasi reliabilitas, daya pembeda soal, tingkat kesukaran, korelasi skor

butir dengan total, kualitas pengecoh atau distraktor, dan kemampuan penalaran statistik mahasiswa terhadap instrumen tes yang dikembangkan.

a. Estimasi Validitas

Validasi bertujuan untuk mengetahui sejauh mana ketepatan dan kecermatan suatu alat ukur dalam melakukan fungsi ukurnya. Adapun dalam penelitian ini validitas yang digunakan adalah validitas isi (*conten validity*). Setelah penyusunan instrumen tes yang berupa pilihan ganda, maka instrumen tes tersebut dan perangkat yang telah dibuat diajukan kepada tim ahli (*expert judgement*) untuk dianalisis secara teoritik dan mengetahui validitas pada instrumen tes yang akan dikembangkan ini. Validasi isi dalam penelitian ini dilakukan oleh tim ahli yang terdiri dari 8 orang. Hasil validitas isi dengan menggunakan CVR dapat dilihat pada tabel berikut :

Tabel 1
Rekapitulasi Hasil Validitas Isi I dengan CVR

Butir Soal	CVR	CVR minimum	Keterangan
1	0.5	0.75	Butir tidak mendukung validitas isi tes
2	0.75		Butir mendukung validitas isi tes
3	0.75		Butir mendukung validitas isi tes
4	0.5		Butir tidak mendukung validitas isi tes
5	0.5		Butir tidak mendukung validitas isi tes
6	0.75		Butir mendukung validitas isi tes
7	0.75		Butir mendukung validitas isi tes
8	0.75		Butir mendukung validitas isi tes
9	0.75		Butir mendukung validitas isi tes
10	0.5		Butir tidak mendukung validitas isi tes
11	0.75		Butir mendukung validitas isi tes
12	0.75		Butir mendukung validitas isi tes
13	0.75		Butir mendukung validitas isi tes
14	-0.5		Butir tidak mendukung validitas isi tes
15	0.75		Butir mendukung validitas isi tes
16	0.5		Butir tidak mendukung validitas isi tes
17	1		Butir mendukung validitas isi tes
18	1		Butir mendukung validitas isi tes
19	0.75		Butir mendukung validitas isi tes
20	0.75		Butir mendukung validitas isi tes

Berdasarkan tabel 4.1 menunjukkan bahwa terdapat 6 butir soal yang harus direvisi dikarenakan nilai CVR yang dihasilkan CVR minimum. Butir soal tersebut adalah butir soal nomor 1, 4, 5, 10, 14 dan 16. Setelah butir-butir soal yang kurang baik direvisi sesuai masukan tim ahli dan disusun kembali susunan butir soalnya maka kembali dilakukan validasi isi kepada ahli-ahli yang sebelumnya memberikan penilaian. Adapun hasil perhitungan validasi isi yang kedua menggunakan CVR dapat dilihat lebih lengkap dapat dilihat pada tabel 2.

Tabel 2
Rekapitulasi Hasil Validitas Isi II dengan CVR

Butir Soal	CVR	CVR minimum	Keterangan
1	0.75	0.75	Butir mendukung validitas isi tes
2	0.75		Butir mendukung validitas isi tes
3	1		Butir mendukung validitas isi tes
4	1		Butir mendukung validitas isi tes
5	0.75		Butir mendukung validitas isi tes
6	1		Butir mendukung validitas isi tes
7	1		Butir mendukung validitas isi tes
8	0.75		Butir mendukung validitas isi tes
9	0.75		Butir mendukung validitas isi tes
10	1		Butir mendukung validitas isi tes
11	1		Butir mendukung validitas isi tes
12	1		Butir mendukung validitas isi tes
13	1		Butir mendukung validitas isi tes
14	1		Butir mendukung validitas isi tes
15	1		Butir mendukung validitas isi tes
16	1		Butir mendukung validitas isi tes
17	1		Butir mendukung validitas isi tes
18	1		Butir mendukung validitas isi tes
19	1		Butir mendukung validitas isi tes
20	1		Butir mendukung validitas isi tes

Hasil validasi isi pada tabel 3 menunjukkan bahwa dari 20 butir soal yang ditelaah oleh 8 panelis/ahli telah menunjukkan butir-butir tersebut mendukung validitas isi tes. Kemudian dari hasil CVR tersebut maka dihasilkan nilai CVI (*Content Validity Index*) yang merupakan rata-rata dari CVR semua item sebesar 0,94.

b. **Analisis Hasil Uji Coba Terbatas**

Tahap uji coba terbatas dalam penelitian ini dilakukan pada mahasiswa yang berada pada semester II, IV, dan VI jurusan Tadris Matematika IAIN Syekh Nurjati Cirebon dengan jumlah 30 mahasiswa. Tujuan dari tahap uji coba terbatas yaitu untuk mengestimasi reliabilitas, daya pembeda, tingkat kesukaran, korelasi skor butir dengan skor total, kualitas pengecoh atau distraktor butir soal, serta kemampuan penalaran statistik mahasiswa. Instrumen tes yang digunakan dalam penelitian ini sebanyak 20 butir yang telah diperbaiki dari pada tahap validasi isi.

1) **Estimasi Reliabilitas**

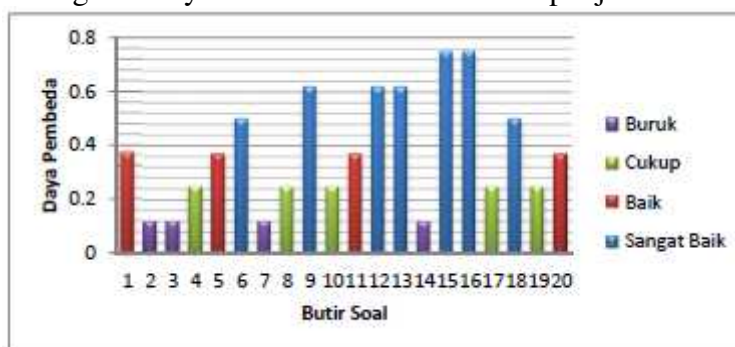
Reliabilitas terkait dengan ketetapan suatu tes apabila diteskan terhadap kelompok yang sama meskipun dilakukan oleh orang yang berbeda (Arikunto, 2012:74). Nilai reliabilitas diperoleh diperoleh dengan metode belah dua atau *split-half method* yaitu dengan cara membelah atas item-item genap dan ganjil. Berdasarkan perhitungan yang dilakukan dengan bantuan software *Anates*, nilai reliabilitas yang dihasilkan adalah sebesar **0,60**. Menurut Arifin (2012:257) nilai reliabilitas antara 0,41 – 0,60 termasuk kategori cukup, dengan kategori cukup berdasarkan kategori yang dikemukakan oleh Arifin (2012:257). Berdasarkan angka estimasi reliabilitas tersebut berarti soal yang

sudah dikembangkan cukup reliabel. Artinya tes yang dikembangkan dapat memberikan hasil yang sama bila diberikan pada kelompok yang sama meskipun dilakukan oleh orang yang berbeda, waktu atau kesempatan yang berbeda dan tempat yang berbeda pula.

2) Daya Pembeda

Analisis daya pembeda dilakukan untuk mengetahui kemampuan suatu soal untuk membedakan antara siswa yang berkemampuan tinggi dengan siswa yang berkemampuan rendah (Arikunto, 2012:226). Pada perhitungan daya pembeda tahap uji coba terbatas, peserta tes dibagi menjadi dua kelompok yaitu mahasiswa kelompok berkemampuan tinggi (kelompok atas) dan mahasiswa kelompok berkemampuan rendah (kelompok bawah). Jumlah peserta tes dalam satu kelompok diambil dari 27% dari jumlah total peserta tes. Sehingga masing-masing kelompok berjumlah 8 peserta tes.

Grafik 1
Diagram Daya Pembeda Butir Soal Tahap Uji Terbatas



Penafsiran daya pembeda didasarkan pada pedoman penafsiran yang dikemukakan oleh Arifin (2012). Butir soal yang memiliki kategori daya pembeda sangat baik memiliki nilai daya beda pada rentang 0.40 – 1.00. Butir soal yang memiliki kategori tersebut adalah soal nomor 6, 9, 12, 13, 15, 16 dan 18. Butir soal yang memiliki kategori daya pembeda baik memiliki nilai daya beda pada rentang 0.30 – 0.39. Butir soal yang memiliki kategori tersebut adalah soal nomor 1, 5, 11 dan 20. Butir soal yang memiliki kategori daya pembeda cukup memiliki nilai daya beda pada rentang 0.20 – 0.29. Butir soal yang memiliki kategori tersebut adalah soal nomor 4, 8, 10, 17 dan 19. Butir soal yang memiliki kategori daya pembeda buruk memiliki nilai daya beda pada rentang 0.00 – 0.19. Butir soal yang memiliki kategori tersebut adalah soal nomor 2, 3, 7 dan 14.

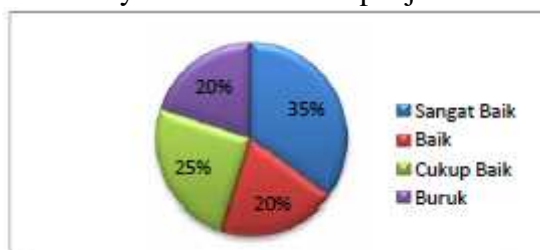
Hasil analisis daya pembeda dalam pengembangan instrumen tes untuk mengukur kemampuan penalaran statistik mahasiswa menunjukkan beberapa kategori yaitu

buruk, cukup, baik dan sangat baik. Dari kategori-kategori tersebut hasil analisisnya menunjukkan bahwasannya butir soal yang dikembangkan dalam penelitian ini lebih banyak yang berkategori sangat baik dengan jumlah 7 butir soal dan yang paling sedikit yaitu pada dua kategori yakni baik dan jelek. Pada daya pembeda kategori baik berjumlah 4 butir soal dan kategori jelek 4 butir soal.

Adapun salah satu butir soal yang berkategori sangat baik yang merupakan jumlah butir soal tertinggi diantara kategori yang lain adalah butir soal 15 (dapat dilihat pada lampiran A.6). Butir soal ini dikatakan sangat baik karena dapat membedakan antara kelompok mahasiswa berkemampuan tinggi (kelompok atas) dengan kelompok mahasiswa berkemampuan rendah (kelompok bawah). Pada butir ini mahasiswa yang dianggap berkemampuan tinggi (kelompok atas) yang menjawab benar sebanyak 6 peserta tes, sedangkan mahasiswa yang dianggap berkemampuan rendah (kelompok bawah) tidak ada yang menjawab benar. Dengan demikian selisih antara kelompok atas dengan bawah adalah 6 dan indeks daya bedanya sebesar 0,75 yang menunjukkan butir tersebut termasuk kategori yang sangat baik.

Kemudian salah satu butir soal yang berkategori buruk adalah butir soal 7 (dapat dilihat pada lampiran A.6). Pada butir ini mahasiswa yang dianggap berkemampuan tinggi (kelompok atas) yang menjawab benar sebanyak 3 peserta tes, sedangkan mahasiswa yang dianggap berkemampuan rendah (kelompok bawah) sebanyak 2 peserta tes. Dengan demikian selisih antara kelompok atas dengan bawah adalah 1 dan indeks daya bedanya sebesar 0,12 yang menunjukkan butir tersebut termasuk kategori yang buruk, sehingga kemampuan membedakan antara kelompok mahasiswa berkemampuan tinggi (kelompok atas) dengan kelompok mahasiswa berkemampuan rendah (kelompok bawah) tidak terlalu baik. Persentase kategori daya pembeda pada tes secara lengkap dapat dilihat pada diagram berikut :

Diagram 1
Persentase Daya Beda Soal Tahap Uji Coba Terbatas



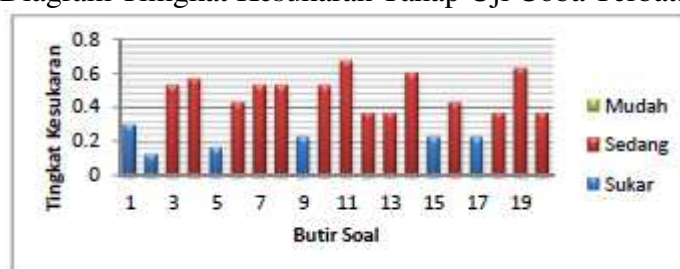
Berdasarkan gambar 4.2 sebanyak 20% dari instrumen tes yang dikembangkan termasuk kategori buruk. Artinya kemampuan butir-butir tersebut masih kurang dalam membedakan kemampuan mahasiswa berkemampuan tinggi dengan mahasiswa

berkemampuan rendah, sehingga masih memerlukan perbaikan untuk menjadi butir soal yang layak digunakan.

3) Tingkat Kesukaran

Soal yang baik adalah soal yang memiliki tingkat kesukaran yang seimbang (Arifin, 2012). Oleh karena itu dalam penyusunan instrumen tes ini perlu memperhatikan tingkat kesukarannya. Berikut adalah diagram hasil perhitungan tingkat kesukaran tiap butir soal :

Grafik 2
Diagram Tingkat Kesukaran Tahap Uji Coba Terbatas



Berdasarkan gambar 4.3 bahwa ada beberapa kategori tingkat kesukaran pada tahap uji coba terbatas yang dibagi ke dalam tiga kategori yaitu soal yang tergolong mudah, sedang dan sukar (DEPDIKNAS, 2008). Kategori mudah memiliki nilai tingkat kesukaran 0.71 – 1.00. Butir soal yang tergolong kategori mudah tidak ada. Kategori sedang memiliki nilai tingkat kesukaran 0.31 – 0.70. Butir soal yang tergolong kategori sedang adalah butir soal 3, 4, 6, 7, 8, 10, 11, 12, 13, 14, 16, 18, dan 20. Sementara kategori sukar memiliki nilai tingkat kesukaran 0.00 – 0.30. Butir soal yang tergolong kategori sukar adalah 1, 2, 5, 9, 15 dan 17.

Hasil analisis tingkat kesukaran soal dalam pengembangan instrumen tes untuk mengukur kemampuan penalaran statistik mahasiswa adalah soal-soal tersebut lebih banyak termasuk pada kategori soal yang sedang dan itu menunjukkan soal-soal tersebut baik. Sedangkan pada tahap uji coba terbatas ini tidak ada yang berkategori mudah.

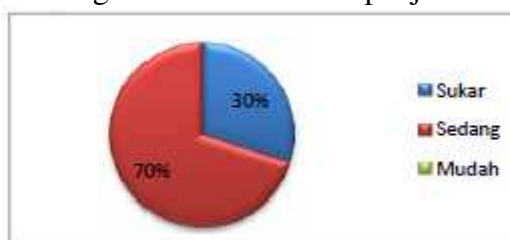
Untuk kategori sedang yang merupakan kategori dengan jumlah butir soal yang paling tinggi yaitu berjumlah 14 butir. Salah satu kutipan soal yang berkategori sedang yaitu soal bernomor 3 (dapat dilihat pada lampiran A.6). Butir soal tersebut termasuk pada indikator mampu menentukan hubungan pada ukuran pemusatan dan soal tersebut mampu dijawab dengan benar oleh 16 peserta tes. Adapun indeks tingkat kesukarannya adalah 53,33% atau 0,53 sehingga soal tersebut dikategorikan sedang.

Untuk kategori sukar terdapat 6 butir soal, salah satunya adalah butir soal bernomor 15 (dapat dilihat pada lampiran A.6). Butir soal tersebut termasuk indikator mampu

menentukan variabilitas sampling dan soal tersebut mampu dijawab dengan benar oleh 7 peserta tes. Adapun indeks tingkat kesukarannya adalah 23,33% atau 0,23 sehingga soal tersebut dikategorikan sukar.

Berdasarkan penjelasan di atas maka tingkat kesukaran itu dapat dilihat dari jumlah peserta tes yang menjawab benar butir soal berbanding jumlah siswa yang mengikuti tes. Semakin banyak yang menjawab suatu soal, maka soal tersebut semakin dikategorikan mudah, sebaliknya semakin sedikit suatu soal di jawab maka soal tersebut semakin dikategorikan sukar.

Diagram 2
Persentase Tingkat Kesukaran Tahap Uji Coba Terbatas



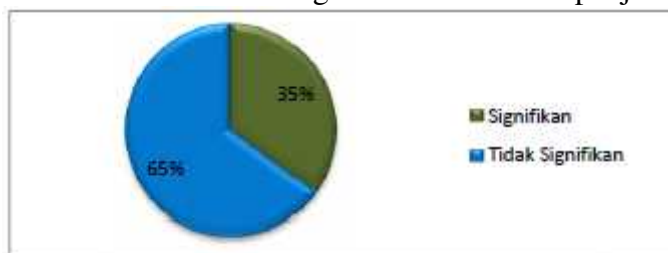
Berdasarkan diagram 2 dapat diketahui butir soal pada tahap uji coba terbatas lebih banyak berkategori sedang yaitu sebesar 70%. Artinya hampir 70% peserta tes pada tahap ini dapat menjawab butir-butir soal dengan benar.

4) Korelasi Skor Butir dengan Skor Total

Korelasi skor butir dengan skor total menggambarkan skor yang didapatkan pada butir tersebut mempengaruhi dan mempunyai hubungan dengan skor secara keseluruhan (skor total). Dalam korelasi ini terbagi menjadi dua kategori yaitu butir soal yang signifikan dan tidak signifikan. Untuk butir soal yang signifikan artinya skor butir soal tersebut secara signifikan memang berkorelasi dengan skor total yang ada. Sementara butir soal yang tidak signifikan artinya tidak ada korelasi yang signifikan antara skor butir tersebut dengan skor total.

Sebuah butir soal dinyatakan memiliki korelasi yang signifikan jika mempunyai harga korelasi r_{pb} r tabel (Sudijono,2012:259). Dengan demikian, jika butir soal tidak memenuhi kriteria tersebut maka butir tersebut tidak signifikan.

Diagram 3
Persentase Korelasi Skor Butir dengan Skor Total Tahap Uji Coba Terbatas



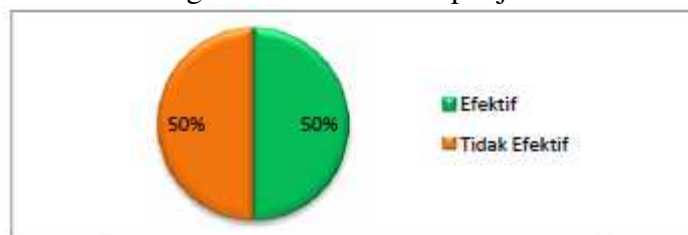
Berdasarkan diagram 3, bahwa analisis korelasi skor butir dengan skor total pada pengembangan instrumen tes pada tahap uji coba terbatas lebih banyak kategori yang tidak signifikan yaitu dengan jumlah 13 butir dengan persentase sebesar 65% dibandingkan dengan kategori yang signifikan. Artinya butir soal yang termasuk kategori tidak signifikan itu tidak mempengaruhi skor total yang ada.

5) Kualitas Distraktor

Kualitas distraktor dapat dikatakan berfungsi apabila paling tidak dipilih oleh 5% peserta tes dan juga lebih banyak dipilih oleh kelompok mahasiswa yang belum paham materi atau kelompok bawah (DEPDIKNAS, 2008:15). Dalam penelitian ini peserta tes yang mengikuti tes pada tahap uji coba terbatas sebanyak 30 orang.

Jadi, distraktor (pengecoh) dikatakan efektif apabila pilihan jawaban dapat dipilih minimal oleh 2 peserta tes. Sedangkan distraktor (pengecoh) dikatakan tidak efektif apabila pilihan jawaban dapat dipilih kurang dari 2 peserta tes.

Diagram 3
Persentase Tingkat Distraktor Tahap Uji Coba Terbatas



Dilihat dari diagram 3, bahwa tingkat distraktor terbagi menjadi dua kategori yaitu efektif dan tidak efektif. Adapun hasil analisis tingkat distraktor butir soal pada tahap uji coba terbatas dalam penelitian ini seimbang antara kategori yang efektif dan tidak efektif dengan persentase masing-masing kategori 50%. Hal ini menunjukkan bahwa distraktor (pengecoh) cukup berfungsi mengecoh jawaban peserta tes, karena peserta tes pada tahap uji coba terbatas pada kelompok mahasiswa yang berkemampuan rendah memilih jawaban yang salah sehingga terkecoh dengan jawaban yang benar.

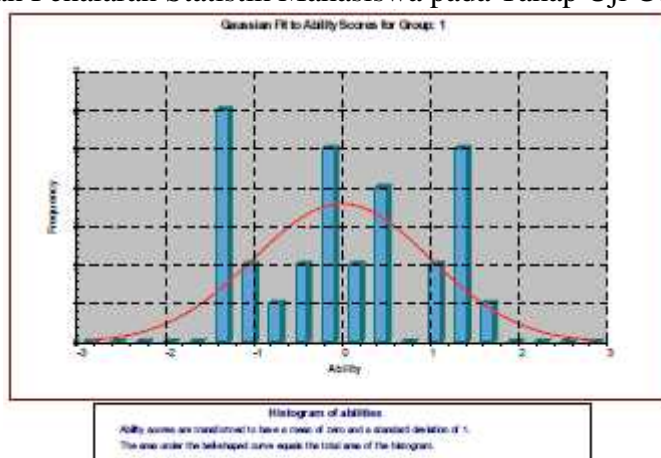
6) Estimasi Kemampuan Penalaran Statistik

Kemampuan penalaran statistik masing-masing peserta tes dilakukan dengan menggunakan perhitungan IRT model satu parameter dengan bantuan software *BILOG-MG 3.00* pada phase 3. Kategori nilai estimasi kemampuan/*ability* tergolong menjadi empat kategori yaitu kategori kurang, rata-rata (cukup), baik dan baik sekali (Erawan, 2009). Kategori “kurang” untuk nilai estimasi kemampuan lebih kecil dari -1.5, kategori rata-rata (cukup) untuk nilai estimasi kemampuan antara -1.5 sampai dengan kurang dari 1.5, kategori “baik” untuk nilai estimasi kemampuan antara 1.5

sampai 3.00, dan kategori “baik sekali” untuk nilai estimasi kemampuan lebih dari 3.00.

Berdasarkan data hasil out put BILOG MG-3.00 pada phase 3 (dapat dilihat pada lampiran), kemampuan peserta tes pada tahap uji coba terbatas yang berjumlah 30 peserta tes semuanya memiliki nilai *ability*/kemampuan antara -1.5 sampai dengan kurang dari 1.5. Dengan demikian hasil estimasi kemampuan dari 30 peserta tersebut berada pada kategori rata-rata atau cukup, sementara kategori lainnya tidak ada. Berikut diagram kemampuan penalaran statistik mahasiswa pada tahap uji coba terbatas.

Grafik 3
Kemampuan Penalaran Statistik Mahasiswa pada Tahap Uji Coba Terbatas



Selain mengetahui hasil estimasi kemampuan penalaran statistik masing-masing peserta tes, pada tahap uji coba terbatas ini juga mengetahui indikator kemampuan penalaran statistik manakah yang dominan pada indikator tes yang dikembangkan.

INTERPRETASI DAN PEMBAHASAN

1. Kualitas Instrumen Tes yang Dikembangkan

Berdasarkan hasil yang diperoleh dari uji coba empirik, dilakukan analisis butir soal pada tiap tahap dalam penelitian pengembangan instrumen tes untuk mengukur kemampuan penalaran statistic mahasiswa. Analisis tersebut meliputi estimasi validitas isi, reliabilitas, daya pembeda soal, tingkat kesukaran, korelasi skor butir dengan skor total, dan tingkat distraktor soal.

Hasilnya ketika instrumen tes yang telah divalidasi oleh ahli maka diberikan kepada peserta uji coba terbatas kemudian dianalisis dan dilakukan revisi untuk digunakan pada tahap uji coba luas. Instrumen tes yang telah dikembangkan dalam penelitian ini secara keseluruhan memiliki kualitas yang baik dari segi validitas, reliabilitas, daya pembeda, tingkat kesukaran, serta tingkat distraktornya. Akan tetapi dari segi korelasi skor butir dengan total kurang baik. Hasil analisis kualitas instrumen tes yang dikembangkan dapat dijelaskan sebagai berikut:

Hasil validasi isi yang dilakukan oleh para ahli menyatakan bahwa instrumen tes yang dikembangkan memiliki CVI (*Content Validity Index*) sebesar 0,94. Menurut Lawshe (1975) artinya butir soal yang terdapat dalam instrumen tersebut dapat mencakup keseluruhan kawasan isi tes yang mencerminkan kemampuan penalaran statistik.

Berdasarkan hasil uji coba empirik, koefisien reliabilitas hasil estimasi adalah 0,68 dengan kategori tinggi. Berdasarkan koefisien reliabilitas tersebut berarti soal yang dikembangkan dikatakan reliabel dengan kualitas tinggi. Artinya tes yang dikembangkan dapat memberikan hasil yang sama bila diberikan pada kelompok yang sama meskipun dilakukan oleh orang yang berbeda, waktu atau kesempatan yang berbeda dan tempat yang berbeda pula sehingga konsistensi tes ini dianggap tinggi dan dapat dipercaya (Arifin, 2012:257).

Untuk analisis butir soal yaitu dari segi daya pembeda, kategori daya pembeda soal sangat baik sebanyak 11 butir yaitu butir soal 1, 2, 3, 4, 7, 8, 13, 15, 16, 17 dan 18 dengan persentase sebesar 55%, kategori baik sebanyak 3 butir yaitu butir soal 11, 19, dan 20 dengan persentase sebesar 15%, kategori cukup sebanyak 3 butir yaitu butir soal 10, 12 dan 14 dengan persentase sebesar 15%, dan kategori buruk sejumlah 3 butir yaitu butir 5, 6 dan 9 dengan persentase sebesar 15%. Menurut Arikunto (2012:226) butir soal yang memiliki kualitas daya pembeda yang baik adalah butir soal yang dapat membedakan peserta tes berkemampuan tinggi dengan peserta tes berkemampuan rendah. Butir soal tersebut adalah butir soal yang memiliki kategori sangat baik dan baik.

Berdasarkan analisis butir soal dari segi tingkat kesukaran, diperoleh bahwa instrumen tes yang dikembangkan dalam penelitian ini semuanya berkategori sedang. Dengan demikian dari segi tingkat kesukaran, instrumen tes yang dikembangkan memiliki kualitas yang baik. Hal ini berdasarkan pendapat Nitko (1996), bahwa suatu soal dikatakan efektif tingkat kesukarannya apabila soal tersebut dapat dijawab benar oleh semua kelompok yaitu kelompok atas, menengah maupun bawah. Soal tersebut adalah soal dengan tingkat kesukaran dengan kategori sedang yang dapat dijawab oleh semua kelompok.

Berdasarkan analisis butir soal dari segi korelasi skor butir dengan total, butir soal yang berkategori signifikan berjumlah 12 butir yaitu butir soal 1, 2, 3, 4, 7, 8, 13, 16, 17, 18, 19, dan 20 dengan persentase sebesar 60%. Butir soal yang berkategori tidak signifikan berjumlah 8 butir yaitu butir soal 5, 6, 9, 10, 11, 12, 14, dan 15 dengan persentase sebesar 40% . Menurut Sudijono (2012:259) bahwa butir yang dapat menggambarkan skor yang didapatkan pada butir tersebut mempengaruhi terhadap skor secara keseluruhan adalah butir yang memiliki korelasi yang signifikan.

Berdasarkan analisis butir soal dari segi kualitas distraktor butir soal dalam penelitian ini, lebih banyak pada kategori yang efektif dengan persentase 65%. Butir tersebut adalah 1,2, 5, 6, 8, 9,

10, 12, 13, 15, 16, 17, dan 19. Sisanya yang tidak efektif yaitu butir 3, 4, 7, 11, 14, 18, dan 20. Hal ini menunjukkan bahwa distraktor (pengecoh) dapat berfungsi dengan baik mengecoh jawaban peserta tes, karena lebih banyak yang berkategori efektif yaitu peserta tes pada kelompok yang berkemampuan rendah memilih jawaban yang salah sehingga terkecoh dengan jawaban yang benar. (DEPDIKNAS, 2008:15)

Berdasarkan penjelasan di atas, ternyata tidak semua butir yang menurut para ahli dinyatakan bahwa butir tersebut mendukung validitas isi, akan tetapi dari hasil empirik butir tersebut memiliki kualitas yang buruk dari beberapa kriteria yang dinilai. Seperti butir 14, menurut para ahli butir tersebut mendukung validitas isi, akan tetapi dari hasil uji coba empirik butir tersebut hanya memiliki kualitas baik dari kriteria tingkat kesukaran.

Hasil yang diperoleh dari uji coba empirik semuanya hanyalah angka-angka hasil estimasi yang dipengaruhi oleh beberapa faktor. Pada kriteria tingkat kesukaran, faktor yang menyebabkan butir soal berkategori sukar yaitu :

- a. Butir soal itu "mungkin" salah kunci jawaban.
- b. Butir soal itu mempunyai 2 atau lebih jawaban yang benar.
- c. Materi yang ditanyakan belum diajarkan atau belum tuntas pembelajarannya, sehingga kompetensi minimum yang harus dikuasai peserta didik belum tercapai.
- d. Pernyataan atau kalimat soal terlalu kompleks dan panjang.

Sementara itu, pada kriteria daya pembeda apabila suatu butir soal tidak dapat membedakan kedua kemampuan peserta itu, maka butir soal itu dapat dicurigai "kemungkinannya" seperti berikut ini:

- a. Kunci jawaban butir soal itu tidak tepat.
- b. Butir soal itu memiliki 2 atau lebih kunci jawaban yang benar
- c. Pengecoh tidak berfungsi
- d. Materi yang ditanyakan terlalu sulit, sehingga banyak peserta tes yang menebak
- e. Sebagian besar peserta yang memahami materi yang ditanyakan, berpikir ada yang salah informasi dalam butir soalnya.

Walaupun terdapat beberapa faktor yang mempengaruhi angka-angka estimasi yang dihasilkan pada tahap pengembangan instrumen tes dalam penelitian ini, diharapkan angka-angka hasil estimasi tersebut lebih besar dipengaruhi oleh pengembangan yang dilakukan dalam penelitian instrumen tes ini.

2. Hasil Analisis Pengukuran Kemampuan Penalaran Statistik dengan Menggunakan Instrumen Tes yang Telah Dikembangkan

Berdasarkan hasil uji coba empirik, selain diperoleh kualitas tes yang dikembangkan juga diperoleh ukuran kemampuan penalaran statistic peserta tes. Kemampuan ini dianalisis menggunakan software *BILOGMG 3.00*. Analisis ini bertujuan untuk mengetahui nilai estimasi kemampuan/*ability* peserta yang dikategorikan dalam empat kategori yaitu kurang, cukup, baik, dan baik sekali.

Berdasarkan hasil analisis diperoleh bahwa hampir keseluruhan kemampuan penalaran statistik mahasiswa dengan menggunakan instrumen tes yang telah dikembangkan memiliki nilai estimasi kemampuan/*ability* dengan kategori cukup yaitu antara -1,5 sampai dengan kurang dari 1,5 dengan persentase sebesar 99% dari seluruh peserta tes. Sementara sisanya hanya 1% peserta tes yang memiliki nilai estimasi kemampuan/*ability* dengan kategori baik yaitu antara 1,5 sampai dengan kurang dari 3.

KESIMPULAN

Berdasarkan data yang dihasilkan dari penelitian yang telah dilakukan, dengan mengacu pada rumusan masalah penelitian yang telah dibuat sebelumnya maka dapat diambil kesimpulan sebagai berikut:

1. Instrumen tes yang dikembangkan secara keseluruhan termasuk dalam kualitas yang baik dilihat dari validitas, reliabilitas, daya pembeda serta tingkat distrakturnya. Tes yang dikembangkan memiliki indeks validitas isi (CVI) sebesar 0,94 dan reliabilitas dengan kategori tinggi sebesar 0,68.
2. Hasil analisis butir soal terhadap 20 butir soal didapatkan bahwa butir soal yang memiliki kategori daya pembeda soal sangat baik sebanyak 11 butir kategori daya pembeda soal baik sebanyak 3 butir, kategori daya pembeda soal cukup sebanyak 3 butir, dan kategori daya pembeda soal buruk sebanyak 3 butir. Tes ini memiliki tingkat kesukaran satu kategori yaitu kategori sedang untuk semua butir soal. Korelasi skor butir dengan skor total dari instrumen tes yang dikembangkan ini memiliki kategori signifikan berjumlah 12 butir dan kategori tidak signifikan berjumlah 8 butir. Untuk kualitas distraktor dari instrumen tes yang dikembangkan ini memiliki kategori yang efektif sebanyak 13 butir dan yang memiliki kategori tidak efektif sebanyak 7 butir.
3. Berdasarkan hasil analisis data menggunakan IRT, hampir keseluruhan kemampuan penalaran statistik mahasiswa dengan menggunakan instrumen tes yang telah dikembangkan berkategori cukup. Sebanyak 99% mahasiswa memiliki nilai estimasi kemampuan/*ability* dengan kategori cukup yaitu antara -1,5 sampai dengan kurang dari 1,5, sementara sisanya yaitu 1% memiliki nilai estimasi kemampuan/*ability* dengan kategori baik yaitu antara 1,5 sampai dengan kurang dari 3.

SARAN

1. Mengingat pentingnya kemampuan penalaran statistik bagi mahasiswa, pendidik disarankan untuk menggunakan tes-tes yang mendukung berkembangnya kemampuan penalaran statistik.
2. Sebaiknya dilakukan penelitian pengembangan instrumen tes untuk mengukur kemampuan penalaran statistik kembali, agar diperoleh instrumen tes yang jauh lebih baik.
3. Untuk mengembangkan instrumen tes lebih lanjut, disarankan jumlah soal yang diujicobakan pada tahap uji coba diperbanyak dan disesuaikan dengan waktu dan kondisi peserta tes.

DAFTAR PUSTAKA

1. Anggreyani, Arie. 2009. *Penerapan Teori Uji Klasik dan Teori Respon Butir dalam Mengevaluasi Butir Soal*. Skripsi pada Departemen Statistika IPB: tidak diterbitkan.
2. Arifin, Zaenal. 2012. *Evaluasi Pembelajaran*. Bandung : Rosda Karya.
3. Arikunto, Suharsimi. 2012. *Dasar-dasar Evaluasi Pendidikan*. Jakarta: Bumi Aksara.
4. Azwar , Saifuddin. 2010. *Tes Prestasi*. Yogyakarta: Pustaka Pelajar.
5. Azwar , Saifuddin. 2012. *Reliabilitas dan Validitas*. Yogyakarta: Pustaka Pelajar.
6. Soryosubroto. 2009. *Proses Belajar Mengajar di Sekolah*. Jakarta: Rineka Cipta.
7. DelMas, Robert C. 2002. *Statistical Literacy, Reasoning and Learning*. Journal of Statistics Education.
8. Dewi, Nisa. 2011. *Pengaruh Pendekatan Problem Solving terhadap Kemampuan Penalaran Matematis Siswa*. Skripsi pada jurusan Tadris Matematika IAIN Syekh Nurjati Cirebon : tidak diterbitkan.
9. Emzir. 2010. *Metodologi Penelitian Pendidikan*. Jakarta: Grafindo Persada.
10. Erawan, Lalang dan Stefanus Sentosa. 2009. *Computer Adaptive Test dengan Pendekatan Item Respon Theory Satu Parameter*. Jurnal Teknologi Informasi, Volume 5 Nomor 2, Oktober 2009, ISSN 1414-9999.
11. Gal dan Garfield. 1997. *Curricular Goal and Assesment Challenges in Statistics Education*. Voorburg : IOS Press.
12. Garfield, Joan B. 1999. *Teaching and Assessing Statistical Reasoning*. University of Minnesota.
13. Garfield, Joan B. 2002. *The Challenge of Developing Statistical Reasoning*. Journal of Statistics Education.
14. Garfield, Joan B. 2003. *Assessing Statistical Reasoning*. Statistics Education Research Journal.
15. Hadjar, Ibnu. 1996. *Dasar-dasar Metodologi Penelitian Kuantitatif dalam Pendidikan*. Jakarta: Raja Grafindo Persada.

16. Hambleton, R.K. (1993). *Principles and selected applications of item response theory*. Dalam R. L. Linn (Eds.), *Educational Measurement* (3rd ed., pp.147-200). Phoenix, AZ: American Council on Education and the Oryx Press.
17. Indrakusuma, Amier Daien. 1966. *Evaluasi Pendidikan Penilaian Hasil Belajar*, tidak diterbitkan.
18. Kerlinger, Fred N. 1990. *Asas-asas Penelitian Behavioral*. Yogyakarta: Gadjah Mada University Press.
19. Lawshe. 1975. *A Quantitative Approach to Content Validity*. Purdue University. *Jurnal Personnel Psychology* 28, 563-575.
20. Lubis, Mawardi. 2008. *Evaluasi Pendidikan Nilai*. Yogyakarta : Pustaka Pelajar.
21. Martadiputra, Bambang Aviv. 2010. *Kajian tentang Kemampuan Melek Statistis (Statistical Literacy), Penalaran Statistis (Statistical Reasoning) dan Berpikir Statistis (Statistical Thinking) Guru SMP/SMA*. Jurnal UPI.
22. Moore, Betsy dan Stanley Tood. 2010. *Critical and Formative Thinking*
23. *Assesment*. -----
24. Mulyadi. 2010. *Pengembangan Model Evaluasi Pendidikan Agama di Sekolah*. Malang: UIN-MALIKI Press.
- b. Nitko, J Anthony. 1996. *Educational Assesment of Student*. Columbus : Pamela Bennet.
25. Putra, Nusa. 2011. *Research & Development Penelitian dan Pengembangan: Suatu Pengantar*. Jakarta : Rajawali Press.
26. Riduwan. 2011. *Belajar Mudah Penelitian*. Alfabeta : Bandung.
27. Santrock, John W. 2007. *Psikologi Pendidikan Edisi Kedua*. Jakarta : Kencana
28. Setyosari, Punaji. 2012. *Metode Penelitian Pendidikan dan Pengembangan*. Kencana : Jakarta.
29. Sevilla, Consuello dkk. 2006. *Pengantar Metode Penelitian*. Jakarta : UI-Press
30. Sudijono, Anas. 2008. *Pengantar Evaluasi Pendidikan*. Jakarta : Raja Grafindo Persada.
31. Sukardi. 2011. *Evaluasi Pendidikan: Prinsip dan Operasionalnya*. Jakarta : Bumi Aksara.
32. Sukmadinta, Nana Syaodih. 2012. *Metodologi Penelitian Pendidikan*. Bandung : Rosda Karya.
33. Sugiyono. 2008. *Metode Penelitian Pendidikan Pendekatan Kuantitatif, Kualitatif, dan R & D*. Bandung : Alfabeta.
34. ----- 2003. *Undang-Undang SISDIKNAS No 20 Tahun 2003*. Departemen Pendidikan Nasional.
35. ----- 2008. *Panduan Analisis Butir Soal*. Departemen Pendidikan Nasional.