Available online at http://journal.uny.ac.id/index.php/jrpm

Jurnal Riset Pendidikan Matematika 6 (2), 2019, 165-176

Kualitas butir bank soal statistika (Studi kasus: Instrumen ujian akhir mata kuliah statistika Universitas Terbuka)

Agus Santoso [1] *, Kartianom Kartianom [2], Gulzhaina K. Kassymova [3]

[1] Universitas Terbuka. Jalan Pd. Cabe Raya, Kec. Pamulang, Kota Tangerang Selatan, Banten 15418, Indonesia.
[2] Jurusan Pendidikan Guru Madrasah Ibtidaiyah, IAIN Bone.
Jalan HOS Cokroaminoto, Kab. Bone, Sulawesi Selatan, Indonesia.
[3] Abai Kazakh National Pedagogical University. Dostyk Ave 13, Almaty 050000, Kazakhstan
E-mail: aguss@ecampus.ut.ac.id, Telp: +6281227216125
* Corresponding Author

ARTICLE INFO

Article history
Received: 17 Dec. 2019;
Revised: 31 Dec. 2019;
Accepted: 3 January. 2020

Keywords
teori respon butir, statistika ekonomi, ujian akhir semester, bank soal; item response theory, economic statistic, final semester exam, item banks

ABSTRACT

Penelitian ini bertujuan untuk mendeksripsikan kualitas butir soal ujian akhir semester mata kuliah statistika ekonomi yang dikembangkan oleh Universitas Terbuka (UT) sebagai dasar dalam mengembangkan bank soal yang terkalibrasi menggunakan pendekatan Teori Respons Butir. Penelitian ini merupakan penelitian deskriptif kuantitatif. Sumber data penelitian ini adalah pola jawaban mahasiswa UT yang telah mengikuti ujian akhir semester (UAS) mata kuliah statistika ekonomi selama enam masa ujian, dengan ukuran sampel sebanyak 23334 mahasiswa. Hasil penelitian ini menunjukkan bahwa butir-butir soal ujian akhir semester mata kuliah statistika ekonomi yang dikembangkan UT: (1) terbukti valid secara konstruk, yakni hanya mengukur satu faktor dominan, yaitu kemampuan statistika ekonomi; (2) memiliki kehandalan yang baik dengan nilai koefisien reliabilitas empiris lebih dari 0,70 (koefisien reliabilitas empiris = 0,7335); (3) dari 140 butir soal yang dikalibrasi terdapat 108 butir soal (25 butir soal berkualitas baik atau tanpa revisi dan 83 butir soal berkualitas kurang baik atau perlu revisi) yang layak disimpan dalam bank soal, sedangkan 32 butir soal berkualitas tidak baik; dan (4) mampu memberikan informasi akurat terkait kemampuan statistika ekonomi mahasiswa pada level kemampuan yang tinggi (-1,3 sampai +4,0).

This study aims to determine the quality of final semester test items of economic statistics course that was developed by Universitas Terbuka (UT) as a basis for developing calibrated item banks using Item Response Theory. This research uses a quantitative descriptive approach. The researcher investigates the answer pattern of the final semester exam (UAS) in the economic statistics course during six periods of the final exams. The sample size in this study was 23334 students. The results of this study indicate that the final semester exam items of economic statistics courses developed by UT: (1) proved to construct valid, i.e. only measure one dominant factor, namely the ability of economic statistics; (2) has good reliability with empirical reliability coefficient values more than 0.70 (empirical reliability coefficient = 0.7335); (3) of the 140 items calibrated there are 108 items (25 items of good quality or without revision and 83 items of poor quality or need to be revised) that are worth keeping in the question bank, while 32 items of quality are not good; and (4) able to provide accurate information related to students' economic statistical abilities at a high level of ability (-1.3 to +4.0)

How to Cite: Santoso, A., Kartianom, K., & Kassymova, G. (2019). Kualitas butir bank soal statistika (Studi kasus: Instrumen ujian akhir mata kuliah statistika Universitas Terbuka). *Jurnal Riset Pendidikan Matematika*, 6(2), 165-176. doi:https://doi.org/10.21831/jrpm.v6i2.28900

PENDAHULUAN

Perguruan tinggi merupakan salah satu penghasil tenaga kerja terampil dan tenaga ahli yang berkarakter serta berinovasi yang memiliki daya saing di dalam dan di luar negeri. Untuk mewujudkan hal tersebut maka perguruan tinggi diwajibkan untuk menyelenggarakan pengajaran, penelitian, dan pengabdian kepada masyarakat (Wibawa, 2017). Untuk mengetahui kualitas pengajaran maka perlu dilakukan kegiatan asesmen dan pengukuran (Mardapi, 2012; Retnawati, 2013). Asesmen merupakan kegiatan menafsirkan kegiatan pengukuran, sementara pengukuran merupakan kegiatan memberikan derajat/angka pada gejala yang diamati, misalnya melalui ujian tengah semester (UTS) atau ujian akhir semester (UAS). Hal ini juga dilakukan oleh staf pengajar di Universitas Terbuka (UT) untuk mengetahui kualitas pengajaran yang telah dilakukan selama satu semester.

Statistika Ekonomi merupakan salah satu mata kuliah pada Fakultas Ekonomi UT yang diujikan pada UAS. Dilihat dari substansinya, materi statistika memiliki kerumitan yang tinggi. Kerumitan materi statistika ini tergambarkan pada capaian prestasi belajar peserta tes, baik yang duduk di bangku sekolah tingkat menengah maupun pada tingkat pendidikan tinggi. Pada sekolah tingkat menengah, laporan hasil TIMSS (*Trends in Mathematics and Science Study*), PISA (*Programme for International Assessment of Student*), dan UN (Ujian Nasional) menunjukkan bahwa umumnya siswa mengalami kesulitan dalam menyelesaikan soal statistika (Kartianom & Mardapi, 2018; Mills & Holloway, 2013; Retnawati, 2017). Kesulitan siswa pada tingkat menengah ini akan berdampak pada performan siswa tersebut ketika berhadapan dengan materi statistika saat duduk di bangku perguruan tinggi (Kien-Kheng & Idris, 2010). Pada tingkat perguruan tinggi, mahasiswa banyak mengalami kesalahan dalam keterampilan proses, memahami soal, dan menggunakan notasi (Firmansyah, 2017). Banyaknya mahasiswa yang mengalami kesalahan atau tidak menjawab dengan benar butir soal statistika ekonomi saat UAS merupakan indikasi dari kerumitan materi tersebut. Selain itu, jawaban benar dan salah yang diberikan oleh peserta tes atau mahasiswa juga menggambarkan kualitas suatu instrumen.

Instrumen yang baik akan memberikan nilai informasi yang lebih tinggi dari pada kesalahan pengukurannya (*error*) (Retnawati, 2013). Nilai informasi yang tinggi akan memberikan gambaran hasil pengukuran yang sebenarnya. Agar dapat memberikan informasi yang tinggi, maka instrumen yang akan digunakan pada kegiatan pengukuran haruslah valid dan reliabel (Mardapi, 2012; Retnawati, 2013). Instrumen pengukuran juga memiliki karakteristik yang digambarkan oleh butir soal dari instrumen tersebut. Karakteristik butir ini menggambarkan perilaku butir soal yang direspons oleh peserta tes. Kedua hal tersebut menjadi penting untuk diketahui sebagai bahan untuk perbaikan pengajaran. Oleh karena itu, instrumen yang baik itu dapat digambarkan melalui validitas, reliabilitas, dan karakteristik butirnya.

Dalam kegiatan pengukuran skala luas, validitas dan reliabilitas suatu instrumen merupakan hal yang urgen untuk diketahui. Misalnya, penelitian terkait TIMSS, PISA, dan UN, validitas itu berkaitan dengan kualitas suatu instrumen yang dapat dibuktikan dengan melihat sejauh mana ketepatan instrumen tersebut dalam mengukur apa yang hendak menjadi tujuan ukur (Kartianom & Ndayizeye, 2017; Mardapi, 2012; Muslim et al., 2017; Retnawati, 2013; Rindermann & Baumeister, 2015; Tee & Subramaniam, 2018). Mardapi (2012); Retnawati (2013); dan Wu et al. (2016) bersepakat bahwa validitas suatu instrumen dapat dibuktikan secara isi, konstruk, dan kriteria. Sementara reliabilitas berkaitan dengan keandalan instrumen yang digunakan dalam kegiatan pengukuran dalam menghasilkan informasi atau hasil yang konsisten (Wu et al., 2016). Ukuran keandalan sering dinyatakan dalam indeks antara 0 dan 1, dimana indeks 1 menunjukkan bahwa tes berulang akan memiliki hasil yang identik. Sebaliknya, indeks 0 menunjukkan bahwa skor tes peserta tes dari satu tes ke tes yang lain tidak akan menghasilkan hubungan apa pun. Oleh karena itu, reliabilitas dengan indeks yang tinggi lebih diinginkan karena itu menunjukkan bahwa skor peserta tes pada saat tes dapat "dipercaya".

Selain harus valid dan reliabel, instrumen yang berkualitas tinggi juga harus memperhatikan karakteristik butirnya. Karakteristik butir ini berkaitan dengan perilaku butir yang direspons oleh peserta tes. Respons yang berbeda oleh peserta tes pada butir tertentu secara tidak langsung menggambarkan karakteristik atau perilaku dari butir tersebut, seperti tingkat kesulitan, daya pembeda, efektivitas distraktor, dan tebakan semu (*pseudo guessing*). Instrumen yang baik harusnya tidak didominasi oleh butir yang terlalu mudah dan terlalu sulit. Butir yang terlalu mudah atau terlalu sulit tidak mampu membedakan antara peserta tes yang mampu dan tidak mampu. Namun, bukan berarti butir yang terlalu mudah atau terlalu sulit itu tidak baik atau harus dibuang (tidak digunakan), tergantung instrumen itu digunakan untuk apa. Lain halnya butir dengan daya pembeda yang rendah, butir ini akan memberikan

ISSN 2356-2684 (print), ISSN 2477-1503 (online)

informasi yang tidak akurat (Wu et al., 2016). Sementara efektivitas distraktor berkaitan dengan sejauh mana opsi-opsi selain kunci jawaban memiliki ketertarikan untuk dipilih oleh kelompok dengan kemampuan rendah. Untuk dapat mengetahui karakteristik suatu butir maka perlu dilakukan analisis butir secara empirik (Retnawati, 2016). Analisis butir secara empirik ini dapat didekati secara teori tes klasik dan secara teori respons butir.

Teori tes klasik biasa juga disebut dengan teori skor murni klasik yang didasarkan pada suatu model aditif, yakni skor amatan merupakan penjumlahan dari skor sebenarnya dan skor kesalahan pengukurannya (Crocker & Algina, 2008). Jika dituliskan dalam persamaan matematika maka kalimat tersebut menjadi X = T + E. Pendekatan ini sudah banyak digunakan di hampir semua disiplin ilmu. Penggunaan pendekatan ini tidak lain untuk menaksir parameter kemampuan peserta tes dan parameter butir dari suatu instrumen. Pada pendekatan teori tes klasik, kemampuan siswa dinyatakan dengan skor total yang diperolehnya. Selain itu, pada pendekatan ini parameter butir tergantung pada kemampuan peserta tes (Mardapi, 2012). Dengan kata lain, jika suatu instrumen diteskan pada peserta tes dengan kemampuan tinggi maka parameter kesulitan butir tes dari instrumen tersebut cenderung berkategori mudah, begitu pun sebaliknya. Kesalahan pengukuran dalam pendekatan ini juga berlaku untuk semua peserta tes. Hal ini yang kemudian oleh sebagian ahli dijadikan sebagai bagian dari kelemahan pendekatan teori tes klasik yang dikenal dengan istilah *group dependent* (Retnawati, 2016). Oleh karena itu, hadirlah pendekatan modern sebagai alternatif untuk mengatasi kelemahan dari pendekatan teori tes klasik yang disebut dengan teori respons butir. Pendekatan ini lebih memperhatikan interaksi antara setiap peserta tes dengan butir tes.

Pada teori respons butir unidimensi, karakteristik butir instrumen juga berupa tingkat kesulitan dan daya pembeda, namun cara estimasinya yang berbeda (Retnawati, 2016). Ada dua prinsip yang digunakan pada pendekatan ini, yakni prinsip relativitas dan prinsip probabilitas (Keeves & Alagumalai, 1999). Berdasarkan prinsip ini, kemudian dapat disusun suatu model logistik dengan menghubungkan antara peluang seseorang untuk menjawab benar dengan skala kemampuan (θ), tingkat kesulitan (b), daya pembeda butir (a), dan tebakan semu (*pseudo guessing*) (c) (Hambleton et al., 1991). Hubungan keempat parameter tersebut kemudian dinyatakan dalam suatu model persamaan matematis yang memuat 3 parameter butir (model 3-PL) berupa tingkat kesulitan (b), daya pembeda butir (a), dan tebakan semu (*pseudo guessing*) (c); memuat 2 parameter butir (model 2-PL) berupa tingkat kesulitan (b) dan daya pembeda butir (a); dan memuat 1 parameter (model 1-PL) berupa tingkat kesulitan (b). Pendekatan ini banyak digunakan pada penilaian skala luas seperti TIMSS, PISA, UN, dan penilaian skala luas lainnya untuk mengestimasi parameter kemampuan peserta tes dan parameter butir. Hasil estimasi parameter butir dengan pendekatan ini juga digunakan untuk kebutuhan pengembangan bank soal (Retnawati & Hadi, 2014). Keunggulan dari pendekatan ini karena didasari oleh tiga asumsi mendasar yang harus dipenuhi, yakni unidimensi, invariansi parameter, dan independensi lokal.

Di Indonesia, penelitian yang berkaitan dengan analisis kualitas butir suatu instrumen telah banyak dilakukan, namun hanya berfokus pada penggunaan data sekunder tingkat menengah (TIMSS, PISA, UN). Selain itu, pendekatan analisis butir yang digunakan oleh sebagian peneliti juga masih secara klasik. Untuk pendekatan analisis butir secara modern masih jarang dilakukan, hanya pada disiplin ilmu tertentu (misalnya; pada disiplin ilmu psikologi atau psikometri) dan penilaian skala luas tingkat menengah seperti (TIMSS, PISA, UN). Sementara, informasi yang dihasilkan dari pendekatan teori respons butir ini sangat bermanfaat (akurat) dalam meningkatkan kualitas pengajaran, pengembangan sistem bank soal yang terkalibrasi, dan pemetaan kualitas pendidikan. Seperti hasil laporan dari TIMSS dan PISA yang menggunakan pendekatan teori respons butir untuk mengetahui kemampuan peserta tes di seluruh belahan dunia, sehingga bisa dilakukan perbandingan kualitas pendidikan antar negara. Penelitian yang dilakukan oleh Kartianom dan Mardapi (2018) yang memanfaatkan data UN yang dianalisis dengan pendekatan teori respons butir untuk mengetahui kekuatan dan kelemahan siswa yang informasinya dapat digunakan guru sebagai bahan perbaikan pembelajaran, dan penelitian yang dilakukan oleh Retnawati dan Hadi (2014) yang mengembangkan sistem bank soal daerah yang terkalibrasi pada mata pelajaran matematika dengan memanfaatkan parameter butir yang dihasilkan melalui analisis butir dengan pendekatan teori respons butir.

Universitas Terbuka (UT) adalah salah satu perguruan tinggi negeri di Indonesia yang sudah mengembangkan sistem bank soal sejak tahun 1998. Bank soal UT dikembangkan masih berdasarkan pendekatan teori klasik. Lima tahun terakhir, UT berkeinginan untuk mengembangkan (sistem) bank soal dengan pendekatan teori modern. Beberapa analisis butir mulai dicoba untuk dianalisis dan

ISSN 2356-2684 (print), ISSN 2477-1503 (online)

dikalibrasi menggunakan pendekatan teori modern atau teori respons butir (*Item Response Theory*). Hal ini terlihat dari dikembangkannya instrumen untuk mengukur kemampuan statistika ekonomi mahasiswa. Instrumen statistika ekonomi yang dikembangkan terdiri dari enam paket. Butir tes dari masing-masing instrumen diambil atau dipilih dari bank soal yang sudah di kembangkan UT. Jika kita merujuk pada definisi bank soal seperti yang diungkapkan oleh Retnawati dan Hadi (2014) bahwa selain menyusun butir soal juga perlu dilakukan kalibrasi pada butir-butir soal tersebut untuk mendapatkan kualitas butir soal yang baik. Dengan demikian, apa yang dilakukan UT baru sampai pada tahapan penyusunan butir soal, sehingga masih perlu melakukan tahapan selanjutnya berupa analisis butir soal.

Berdasarkan uraian dari beberapa literatur yang telah dikemukakan sebelumnya, untuk dapat mengetahui penyebab banyaknya mahasiswa yang melakukan kesalahan saat menyelesaikan butir soal ujian akhir semester pada mata kuliah statistika ekonomi, maka perlu dilakukan analisis butir soal sehingga dapat diketahui kualitas instrumen yang digunakan saat tes (UAS). Kualitas instrumen ini berkaitan dengan informasi yang akan menggambarkan kualitas pengajaran dan bahan dalam pengembangan sistem bank soal yang terkalibrasi. Dengan demikian, tujuan dari penelitian ini adalah untuk mengetahui kualitas instrumen UAS pada mata kuliah statistika ekonomi di Universitas Terbuka. Kualitas instrumen UAS tersebut digambarkan secara rinci melalui pembuktian validitas konstruk, besaran koefisien reliabilitas, dan karakteristik butir yang dianalisis dengan pendekatan teori respons butir.

METODE

Jenis penelitian ini adalah penelitian deskriptif kuantitatif. Penelitian ini menggunakan *content analysis* untuk mengambil kesimpulan dengan mengidentifikasi berbagai karakteristik khusus suatu pesan secara objektif, sistematis, dan generalis yang terdapat dalam soal dan lembar pola jawaban (respons) peserta tes. Penelitian ini dilaksanakan di kampus Universitas Terbuka Pusat, yang beralamat di Pondok Cabe, Tangerang Selatan, Tangerang, Indonesia. Pengambilan data dilakukan di Pusat Pengujian UT berupa pola respons peserta tes pada saat UAS untuk mata kuliah statistika ekonomi selama enam masa ujian (2015: semester 1 dan 2; 2016: semester 1 dan 2; dan 2017: semester 1 dan 2). Penelitian dilaksanakan selama empat bulan, mulai September 2018 sampai Desember 2018.

Sumber data penelitian ini adala h mahasiswa UT, khususnya mahasiswa Fakultas Ekonomi yang mengikuti UAS dan mengambil mata kuliah statistika ekonomi. Jumlah mahasiswa yang mengikuti UAS sebanyak 23334 mahasiswa. Data ini merupakan data *expost facto* berupa respons peserta tes yang dikumpulkan dengan teknik dokumentasi.

Instrumen UAS pada mata kuliah statistika ekonomi ini terdiri atas enam paket naskah (2015: paket 1 dan 2; 2016: paket 1 dan 2; dan 2017: paket 1 dan 2), setiap paket berisi 30 butir soal, sehingga seluruhnya berjumlah 180 butir soal. Namun hanya 154 butir soal, yang dianalisis lanjut, karena 26 butir soal memiliki daya pembeda butir yang sangat kurang baik (bernilai negatif). Masing-masing paket soal juga memiliki butir bersama (*anchor item*) yang nantinya akan digunakan sebagai penghubung antar paket soal ketika akan dilakukan analisis butir soal secara bersama-sama.

Ada beberapa tahapan yang dilakukan untuk dapat mengetahui kualitas instrumen UAS pada mata kuliah statistika ekonomi, yakni tahap pemetaan butir bersama (*anchor item*), tahap pembuktian validitas konstruk, tahap estimasi reliabilitas, tahap pemilihan model, tahap pengujian data berdasarkan asumsi teori respons butir, dan tahap estimasi parameter butir. Tahap pemetaan butir bersama (*anchor item*) merupakan tahapan awal dalam rangkaian proses analisis butir. *Pertama*, mengidentifikasi butir-butir bersama yang muncul pada setiap paket naskah. *Kedua*, menyusun data dari keenam paket menjadi satu data gabungan. *Software* bantu yang dapat digunakan adalah *Microsoft Excel*.

Tahap selanjutnya adalah pembuktian validitas konstruk. Pembuktian validitas konstruk ini dilakukan dengan cara *Exploratory Factor Analysis* (EFA). *Pertama*, mendeskripsikan nilai Eigen. Nilai Eigen > 1 menggambarkan faktor yang terbentuk. Jika selisih nilai Eigen faktor 1 dan faktor lainnya yang ikut terbentuk sebesar 4 atau persentase kumulatif nilai Eigen minimal sebesar 20%, maka dapat dikatakan instrumen mengukur satu faktor dominan (unidimensi). *Software* bantu yang dapat digunakan adalah IBM SPSS 25. Kedua, mendeskripsikan *scree plot* berdasarkan curaman yang terbentuk. Jika ada satu curaman ekstrim yang terbentuk dibanding curaman lainnya, maka dapat dikatakan instrumen mengukur satu faktor dominan.

Berikutnya adalah tahap estimasi reliabilitas. Reliabilitas suatu instrumen dapat didekati secara teori klasik maupun secara modern. Dalam penelitian ini instrumen akan di estimasi dengan berdasarkan pendekatan modern, yaitu teori respons butir, yakni reliabilitas empiris. Suatu instrumen dikatakan andal

ISSN 2356-2684 (print), ISSN 2477-1503 (online)

atau reliabel jika koefisien reliabilitas dari instrumen tersebut > 0,70 (Mardapi, 2012). *Software* bantu yang dapat digunakan untuk mengestimasi reliabilitas instrumen dengan pendekatan teori respons butir adalah Bilog-MG.

Tahap pemilihan model pada pendekatan teori respons butir juga merupakan tahapan yang tidak kalah penting. Pada tahapan ini akan dipilih 1 model analisis dari 3 model yang ada pada pendekatan teori respons butir, yakni model 1-Parameter Logistik (PL), model 2-PL, dan model 3-PL. *Pertama*, analisis butir berdasarkan model 1-PL, model 2-PL, dan model 3-PL. *Kedua*, kumpulkan nilai *chi-square* dan signifikansinya yang ada pada hasil analisis dari masing-masing model dalam satu tabel. *Ketiga*, membandingkan nilai signifikansi *chi-square* pada masing-masing model dengan *alpha* = 0,05. Jika nilai signifikansi model < 0,05 maka butir dikatakan cocok (*fit*). *Keempat*, akumulasi banyaknya butir yang *fit* (cocok) pada masing-masing model. Model dengan jumlah butir *fit* terbanyak adalah model analisis yang akan dipilih untuk menganalisis butir soal dan mengestimasi kemampuan peserta tes. *Software* bantu yang dapat digunakan adalah Bilog-MG dan Microsoft Excel.

Tahap pengujian data berdasarkan asumsi teori respons butir. Tahapan ini sangat penting dalam pendekatan teori respons butir. *Pertama*, pengujian asumsi unidimensi dengan cara pada pembuktian validitas konstruk. *Kedua*, pengujian invariansi parameter butir dan parameter kemampuan. Untuk pengujian parameter butir, peserta tes dikelompokkan menjadi dua kelompok (ganjil-genap) dan kemudian dilakukan estimasi parameter butir pada masing-masing kelompok. Begitu juga dengan invariansi parameter kemampuan, hanya bedanya yang dibagi menjadi dua kelompok (ganjil-genap) adalah butir soal, bukan peserta tes. Hasil parameter butir dan kemampuan dibuat menjadi *scater plot* yang kemudian dibandingkan dengan garis linear (garis lurus). Jika titik-titik konvergen mendekati garis lurus maka parameter dikatakan invarian. *Ketiga*, pengujian asumsi independensi lokal akan secara otomatis terpenuhi jika asumsi unidimensi terpenuhi. *Software* bantu yang dapat digunakan adalah IBM SPSS 25 dan Microsoft Excel.

Tahap terakhir adalah mendeskripsikan hasil estimasi parameter butir dan kemampuan berdasarkan kriteria model yang terpilih. Adapun kriteria yang digunakan pada model 1-PL, model 2-PL, dan model 3-PL disajikan pada Tabel 1 (Hambleton et al., 1991).

Tabel 1. Kriteria Model Teori Respons Butir

Model	Kriteria Parameter		
	a_i	b_i	c_i
1-PL	0 sampai +2	-	-
2-PL	0 sampai +2	-2 sampai +2	-
3-PL	0 sampai +2	-2 sampai +2	0 sampai 1/k

Keterangana_i = indeks daya pembeda butir; b_i = indeks tingkat kesulitan butir; c_i = indeks *pseudo guessing* (menebak); dan k = banyaknya pilihan jawaban

HASIL DAN PEMBAHASAN

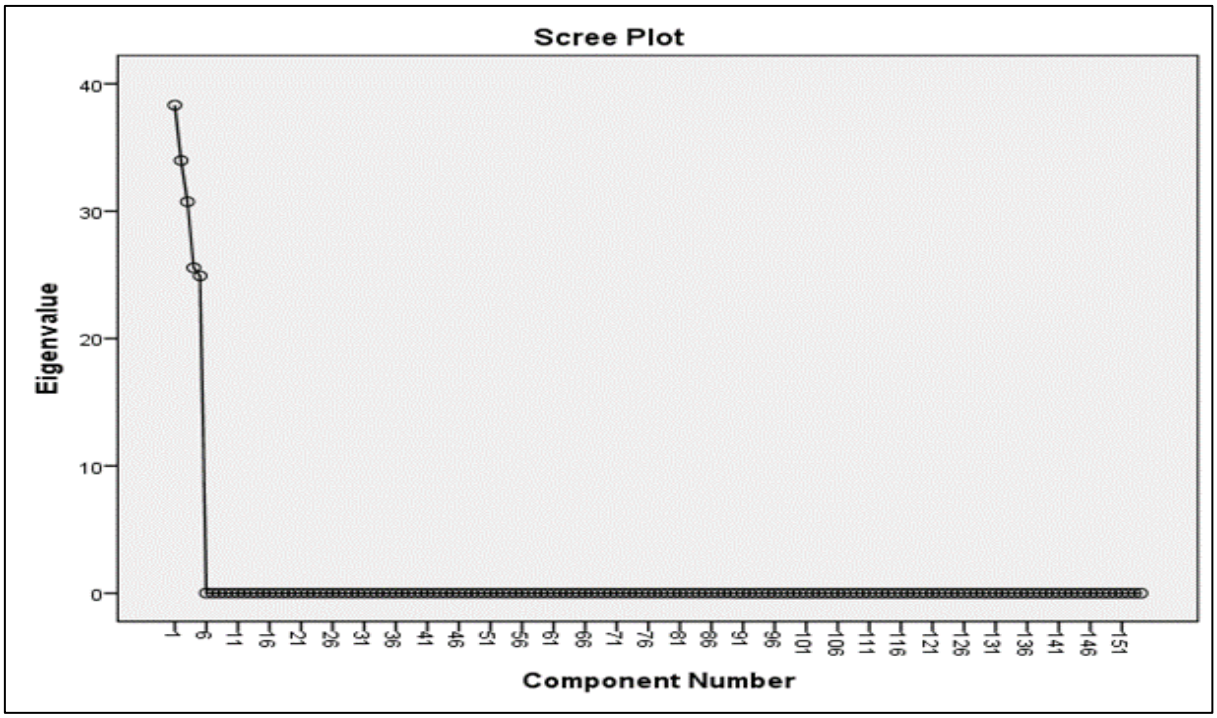
Pembuktian Validitas Instrumen UAS Mata Kuliah Statistika Ekonomi

Instrumen tes yang baik haruslah terbukti valid. Terbukti valid yang dimaksud adalah adanya bukti bahwa instrumen tes yang digunakan memang mengukur apa yang hendak menjadi tujuan ukur. Tujuan dari penggunaan instrumen UAS ini adalah untuk mengukur kemampuan statistika ekonomi mahasiswa UT khususnya di Fakultas Ekonomi UT. Tabel 2 memberikan informasi bahwa hasil analisis faktor menunjukkan adanya faktor dominan. Hal ini terlihat dari persentase kumulatif faktor pertama > 20%.

Tabel 2. Deskripsi Nilai Eigen

Komponen	*Initial Eigenvalues*		
	Total	% *of Variance*	Kumulatif %
1	38.318	24.882	24.882
2	33.974	22.061	46.943
3	30.737	19.959	66.902
4	25.552	16.592	83.494
5	24.916	16.179	99.673
6	0.014	0.009	99.682

ISSN 2356-2684 (print), ISSN 2477-1503 (online)



Gambar 1. *Scree Plot*

Selain berdasarkan informasi pada Tabel 2, informasi berdasarkan *scree plot* pada Gambar 1 juga menunjukkan bahwa hanya terdapat satu faktor dominan. Faktor dominan pada Gambar 1 terlihat dari perubahan nilai eigen dari faktor ke-1 sampai faktor ke-2 yang begitu besar.

Estimasi Reliabilitas Instrumen Statistika Ekonomi

Reliabilitas instrumen UAS statistika ekonomi ini diestimasi dengan metode *likelihood* pada pendekatan teori respons butir. Gambar 2 memberikan informasi bahwa koefisien reliabilitas empiris dari instrumen ini terkategori reliabel dengan besaran > 0,70 atau tepatnya sebesar 0,7335.

```
ROOT-MEAN-SQUARE POSTERIOR STANDARD DEVIATIONS

TEST:              TEST0001
RMS:               0.5179
VARIANCE:          0.2682

EMPIRICAL
  RELIABILITY:     0.7335
```

Gambar 2. Hasil Estimasi Reliabilitas Empiris dengan *Software* Bilog-MG

Karakteristik Instrumen Statistika Ekonomi dengan Pendekatan Teori Respons Butir

Pemilihan Model Analisis Teori Respons Butir

Ada 3 model analisis yang ditawarkan dalam pendekatan teori respons butir, yakni model 1-PL, model 2-PL, dan model 3-PL. Dari ketiga model tersebut akan dipilih satu model analis yang cocok (*fit*) dengan data. Pemilihan model analisis ini dapat dilakukan dengan dua cara, yakni membandingkan nilai signifikansi dari *chi-square* dengan *alpha* = 0,05 dan melihat plot *item information curve* (ICC). Tabel 3 merupakan hasil analisis dari uji kecocokan model berdasarkan nilai *chi-square*. Informasi yang diperoleh dari Tabel 3 bahwa data yang digunakan dalam penelitian ini jika dianalisis dengan pendekatan teori respons butir maka model analisis yang cocok adalah model 3-PL. Hal ini terlihat dari banyak butir yang cocok pada model 3-PL. Dengan demikian, asumsi-asumsi dari teori respons butir yang harus dipenuhi oleh data dan juga dalam mengestimasi parameter harus menggunakan model 3-PL.

ISSN 2356-2684 (print), ISSN 2477-1503 (online)

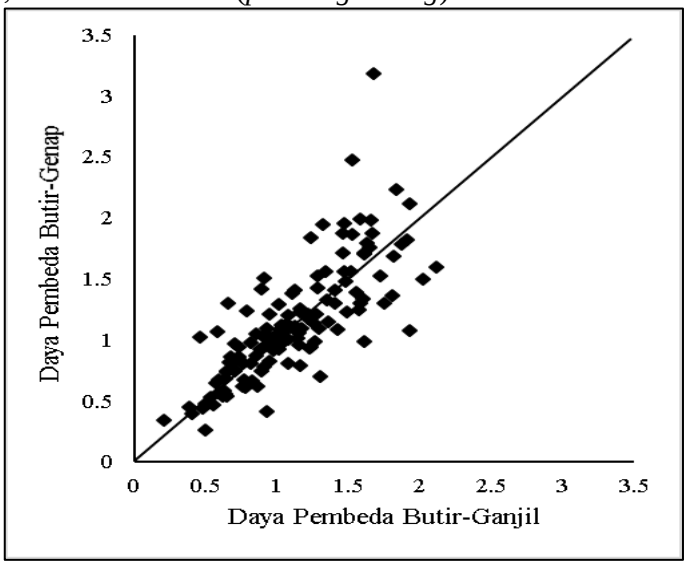
Tabel 3. Hasil Uji Kecocokan Model Analisis Teori Respons Butir

Model Analisis	Jumlah Butir	
	Cocok	Tidak Cocok
Model 1-PL	24	130
Model 2-PL	42	112
Model 3-PL	73	81

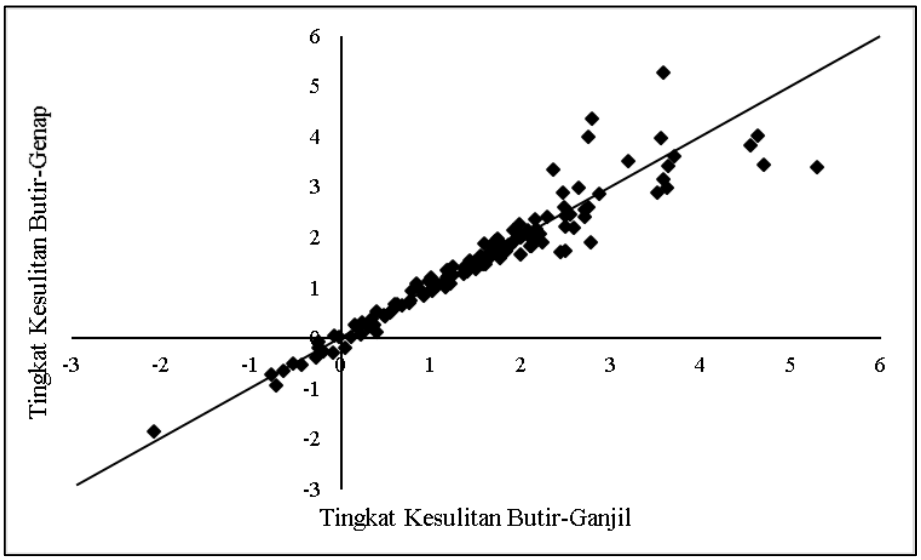
Asumsi Teori Respons Butir

Terbuktinya validitas konstruk instrumen secara tidak langsung juga menjawab asumsi unidimensi pada teori respons butir. Tabel 2 dan Gambar 1 memberikan informasi bahwa instrumen yang digunakan saat UAS mata kuliah Statistika Ekonomi terbukti mengukur satu faktor dominan, yakni kemampuan Statistika Ekonomi mahasiswa. Adanya satu faktor dominan ini dapat pula dikatakan bahwa instrumen tersebut hanya mengukur satu dimensi atau dalam istilah lainnya *unidimensi* (Retnawati, 2016).

Selanjutnya asumsi invariansi parameter butir dapat dilihat dari daya pembeda butir, tingkat kesulitan butir, dan tebakan semu (*pseudo guessing*). Pada Gambar 3, Gambar 4, dan Gambar 5 terlihat titik-titik konvergen mendekati gaaris lurus. Hal ini dapat diartikan bahwa parameter daya pembeda butir, tingkat kesulitan, dan tebakan semu (*pseudo guessing*) bersifat invarian.

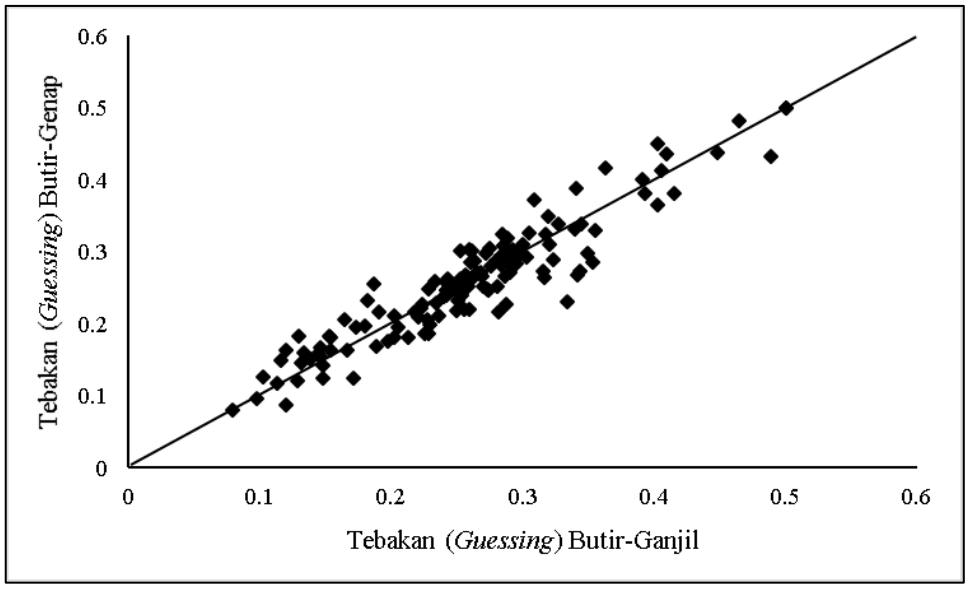


Gambar 3. Parameter Daya Beda



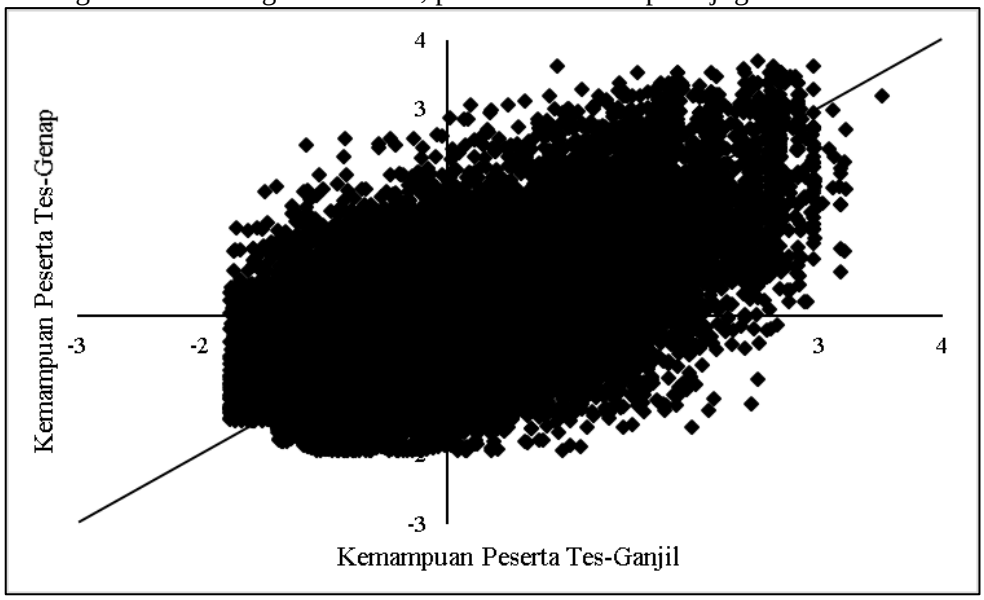
Gambar 4. Parameter Tingkat Kesulitan

ISSN 2356-2684 (print), ISSN 2477-1503 (online)



Gambar 5. Parameter Tebakan Semu

Selain invariansi parameter butir, dalam pendekatan teori respons butir juga menjadi perlu untuk mengetahui invariansi dari parameter kemampuan. Pada Gambar 6 terlihat bahwa titik-titik konvergen juga mendekati garis lurus. Dengan demikian, parameter kemampuan juga bersifat invarian.



Gambar 6. Parameter Kemampuan Karakteristik Butir Instrumen Statistika Ekonomi

Pada Tabel 4 terlihat bahwa dari 154 butir terdapat 14 butir yang tidak ikut terkalibrasi oleh *software* Bilog-MG, sehingga hanya ada 140 butir yang berhasil diestimasi parameternya. Dari 140 butir yang berhasil terkalibrasi, ada 25 butir yang kualitasnya baik, 83 butir kurang baik, dan 32 butir tidak baik.

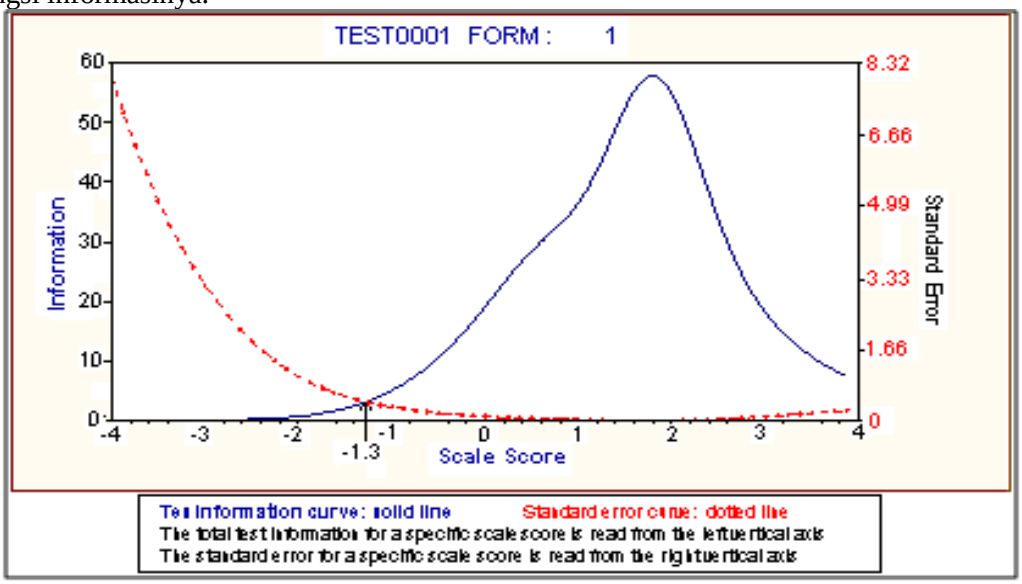
Tabel 4. Karakteristik Butir Soal Berdasarkan Model 3-PL

Kategori	Frekuensi Parameter			Kualitas Butir
	a	b	c	
Baik	134	99	39	25
Kurang Baik	-	-	-	83
Tidak Baik	6	41	101	32
Tidak Terkalibrasi	14	14	14	14
Total	154	154	154	154

ISSN 2356-2684 (print), ISSN 2477-1503 (online)

Fungsi Informasi Tes

Pada Gambar 4 diperoleh informasi bahwa instrumen mata kuliah statistika ekonomi memiliki nilai informasi yang lebih tinggi dibandingkan dengan kesalahan pengukurannya pada rentang kemampuan -1,3 sampai +4,0. Jika soal tersebut diujikan pada peserta tes dengan skala kemampuan kurang dari -1,3 dan lebih dari +4,0, maka yang diperoleh adalah kesalahan pengukuran lebih besar dibandingkan nilai fungsi informasinya.



Gambar 7. Fungsi Informasi dan Kesalahan Pengukuran Tes

Pembahasan

Penelitian ini memberikan informasi bahwa instrumen UAS mata kuliah statistika ekonomi yang dikembangkan Universitas Terbuka telah terbukti valid secara konstruk. Informasi ini berdasarkan hasil analisis faktor secara eksploratori dimana besarnya persentase kumulatif dari nilai Eigen salah satu faktor lebih dari 20% dengan selisih nilai eigen dengan faktor lainnya lebih dari 5 yang digambarkan dengan adanya curaman dominan pada *scree plot*. Kondisi yang demikian menurut Retnawati (2016) adalah bukti bahwa instrumen yang digunakan dalam kegiatan pengukuran valid secara konstruk, yakni hanya mengukur satu faktor dominan atau *unidimensi*. Dengan kata lain, instrumen UAS mata kuliah statistika ekonomi yang dikembangkan terbukti hanya mengukur satu kemampuan, yakni kemampuan statistika ekonomi.

Selain terbukti valid secara konstruk, hasil analisis butir dengan pendekatan teori respons butir juga memberikan informasi bahwa instrumen statistika ekonomi yang dikembangkan oleh Universitas Terbuka ini memiliki keandalan yang baik. Hal ini dilihat dari nilai koefisien reliabilitas empiris instrumen sebesar 0,7335. Mardapi (2012) menjelaskan bahwa instrumen dengan koefisien reliabilitas lebih dari 0,70 memiliki keandalan yang baik. Wu et al. (2016) menambahkan bahwa koefisien reliabilitas yang lebih dari 0,70 menunjukkan skor peserta tes pada saat ujian dapat dipercaya. Dengan kata lain, instrumen UAS mata kuliah statistika ekonomi yang dikembangkan Universitas Terbuka memiliki keandalan yang baik dan informasi yang dihasilkan dapat dipercaya.

Berdasarkan hasil analisis butir dengan pendekatan teori respons butir model 3-PL diperoleh informasi bahwa dari 154 butir instrumen UAS mata kuliah statistika ekonomi, terdapat 25 butir yang berkualitas baik, 83 butir yang berkualitas kurang baik, 32 butir berkualitas tidak baik, dan 14 butir tidak terkalibrasi. Butir yang berkualitas baik ini layak disimpan dalam bank soal karena memiliki daya pembeda, tingkat kesulitan, dan tebakan semu (*pseudo-guessing*) yang baik (Retnawati & Hadi, 2014). Sementara butir yang berkualitas kurang baik masih perlu direvisi lagi sebelum disimpan dalam bank soal. Pada model 3-PL, jika instrumen tes dalam bentuk pilihan ganda maka kemungkinan parameter tebakan semu bernilai tinggi sangat besar (Wu et al., 2016). Solusinya adalah mengkonstruk kembali butir soal atau distraktor butir soal sehingga tidak mudah ditebak oleh peserta tes yang berkemampuan rendah. Sementara, butir soal yang berkualitas tidak baik lebih dikarenakan tidak mampu membedakan

ISSN 2356-2684 (print), ISSN 2477-1503 (online)

kelompok peserta tes yang mampu dan tidak mampu, tingkat kesulitan terlalu tinggi atau terlalu rendah, dan mudah ditebak.

Secara sederhana, daya pembeda butir berkaitan dengan peserta tes mana yang menjawab dengan benar (peserta tes berkemampuan tinggi atau rendah). Dalam teori respons butir, daya pembeda butir yang baik berkisar antara 0 sampai +2 dalam skala l*ogit* (Hambleton et al., 1991). Butir yang memiliki daya pembeda kurang dari 0 maka dapat dikatakan bahwa butir tersebut memiliki daya pembeda yang rendah. Dengan kata lain butir tersebut tidak mampu membedakan antara peserta tes yang mampu dan tidak mampu. Dalam *software* analisis Bilog-MG, butir dengan daya pembeda kurang dari -0,15 maka butir tersebut secara otomatis tidak akan dikalibrasi (Model 2-PL dan Model 3-PL). Wu et al. (2016) menjelaskan bahwa ada tiga alasan yang mungkin terkait daya pembeda butir yang rendah, yakni butir tersebut mengukur hal lain dibandingkan butir-butir yang lainnya (dapat dilihat pada *scree plot* dimana meskipun ada satu curaman dominan namun masih ada empat curaman lainnya yang terbentuk), butir tersebut ditulis atau dikonstruk secara tidak benar sehingga membingungkan peserta tes, dan butir tersebut memiliki tingkat kesulitan yang tinggi (terlalu sulit) atau rendah (terlalu mudah). Wu et al. (2016) menyarankan untuk membuang atau mengganti butir yang memiliki daya pembeda yang rendah karena akan menurunkan reliabilitas tes, meningkatkan kesalahan pengukuran, dan membuat skor tes kurang dapat diartikan atau kurang bermakna.

Sementara tingkat kesulitan butir berkaitan dengan seberapa banyak peserta tes yang menjawab dengan benar. Dalam teori respons butir, tingkat kesulitan butir yang baik berkisar antara -3 sampai +3 dalam skala *logit* (Hambleton et al., 1991). Butir dengan tingkat kesulitan sebesar -3 maka dapat diinterpretasikan bahwa butir tersebut sangat mudah, sedangkan dengan tingkat kesulitan +3 maka dapat diinterpretasikan bahwa butir tersebut sangat sulit. Butir yang sangat mudah atau sangat sulit akan berdampak pada daya pembeda butir yang rendah (Mardapi, 2012). Di sisi lain, Wu et al. (2016) mengatakan bahwa butir yang sangat mudah atau sangat sulit perlu dipertahankan karena instrumen yang digunakan mengukur peserta tes yang memiliki kemampuan rendah dan tinggi. Dengan kata lain, butir dengan tingkat kesulitan sangat mudah bisa ditempatkan pada butir awal untuk mengurangi tingkat kecemasan peserta tes, sementara butir dengan tingkat kesulitan tinggi (dengan catatan masih bisa dijawab benar oleh beberapa peserta tes) akan membantu meningkatkan daya pembeda butir.

Tebakan semu (*pseudo guessing*) merupakan parameter butir yang menggambarkan perilaku menyimpang dari peserta tes dengan kemampuan rendah dalam menjawab butir soal dengan tingkat kesulitan yang lebih tinggi dari kemampuannya secara benar. Barnard-Brak et al. (2018) menambahkan bahwa peserta tes kemampuan rendah yang dimaksud adalah mereka dengan kesempatan belajar (*opportunity to learn*) yang banyak, sementara mereka yang memiliki kesempatan belajar yang sedikit cenderung gagal atau tidak berhasil dalam menebak. Dalam teori respons butir, tebakan semu berkisar antara 0 dan 1 (Hambleton et al., 1991). Hulin et al. (1983) mengatakan bahwa suatu butir dikatakan baik jika memiliki nilai tebakan semu tidak lebih dari 1/k, dimakan k merupakan banyaknya pilihan jawaban (opsi). Rogers (1999) menjelaskan bahwa tebakan semu terjadi pada instrumen yang berbentuk pilihan ganda dan berdampak negatif pada validitas dan reliabilitas instrumen. Attali and Bar-Hillel (2003) menambahkan bahwa antara peserta tes dan pembuat butir soal memiliki kecenderungan yang sama dalam memilih jawaban atau menempatkan kunci jawaban, yakni pada jawaban yang berada di tengah. Temuan dari Attali dan Bar-Hillel (2003) secara tidak langsung menjelaskan maksud tersirat dari pernyataan Barnard-Brak et al. (2018) bahwa kesempatan belajar yang dimaksud salah satunya terkait dengan pola penempatan kunci jawaban pada butir pilihan ganda. Dengan demikian, penyusunan butir soal pilihan ganda secara baik dan benar (khususnya dalam menyusun distraktor) adalah solusi dalam mereduksi terjadinya tebakan semu.

Fungsi informasi tes dan *Standard Error of Measurement* (SEM) merepresentasikan bahwa instrumen UAS mata kuliah statistika ekonomi yang dikembangkan Universitas Terbuka akan memberikan informasi yang akurat jika digunakan untuk mengukur kemampuan statistika ekonomi mahasiswa pada rentang kemampuan -1,3 sampai +4,0. Istiyono et al. (2014) menyatakan bahwa kemampuan pada rentang -0,8 sampai 3,4 termasuk pada level kemampuan yang tinggi. Dengan kata lain, instrumen UAS mata kuliah statistika ekonomi dapat menggali informasi terkait kemampuan statistika ekonomi mahasiswa secara akurat pada mahasiswa dengan level kemampuan yang tinggi. Hal ini terlihat dari nilai informasi tertinggi yang dihasilkan oleh instrumen ini sebesar 58 ketika kemampuan peserta tes atau mahasiswa sebesar 1,75 (termasuk dalam level kemampuan tinggi). Pada kemampuan tersebut kesalahan pengukuran yang dihasilkan hampir mendekati 0 (sangat kecil). Ramos et al. (2013) menjelaskan

ISSN 2356-2684 (print), ISSN 2477-1503 (online)

bahwa peserta tes dengan kemampuan tinggi memiliki pemahaman konsep yang mendalam terhadap suatu materi sehingga akan dapat menggunakan pengetahuan yang dimilikinya secara tepat dalam menyelesaikan masalah yang dihadapinya.

SIMPULAN

Berdasarkan hasil analisis dan pembahasan, dapat disimpulkan beberapa hal. *Pertama*, instrumen UAS mata kuliah statistik ekonomi terbukti valid secara konstruk karena hanya mengukur satu faktor dominan (*unidimensi*), meskipun pada *scree plot* terlihat ada empat faktor lain yang ikut terbentuk. *Kedua*, instrumen UAS mata kuliah statistik ekonomi memiliki keandalan yang baik dengan nilai koefisien reliabilitas empiris lebih dari 0,70. *Ketiga*, dari 140 butir soal instrumen UAS mata kuliah statistika ekonomi yang dikalibrasi terdapat 108 butir (25 butir soal tanpa revisi dan 83 butir soal perlu revisi) yang layak disimpan dalam bank soal, sementara 32 butir soal lainnya perlu diganti atau tidak layak disimpan dalam bank soal karena memiliki daya beda dan tingkat kesulitan tidak baik atau memiliki daya beda dan tebakan semu tidak baik atau memiliki tingkat kesulitan dan tebakan semu tidak baik atau memiliki daya beda, tingkat kesulitan, dan tebakan semu yang tidak baik. *Keempat*, instrumen UAS mata kuliah statistika ekonomi ini akan memberikan informasi atau dapat menggali informasi secara akurat terkait kemampuan statistika ekonomi mahasiswa pada level kemampuan yang tinggi (-1,3 sampai +4).

Temuan penelitian ini memberikan beberapa implikasi bagi praktik maupun penelitian selanjutnya. Pengembang tes hendaknya benar-benar memperhatikan kaidah dalam penyusunan instrumen tes, sehingga dapat dihasilkan butir-butir tes yang berkualitas. Dalam mengembangkan instrumen tes, pengembang soal harus benar-benar ahli di bidangnya serta menguasai kaidah-kaidah dasar dalam menyusun butir tes. Jika diperlukan lembaga-lembaga penyelenggara tes perlu memberikan pelatihan kepada tim penyusun tes terkait kaidah penyusunan tes yang terstandar. Penelitian-penelitian ke depannya hendaknya juga menganalisis kualitas instrumen tes yang dikembangkan oleh lembaga-lembaga lainnya, terutama untuk tes yang dikembangkan secara luas. Penelitian terkait kualitas tes tidak terbatas hanya pada instrumen berbentuk objektif, tetapi juga dapat dilakukan pada instrumen tes berbentuk uraian maupun bentuk tes lainnya.

DAFTAR PUSTAKA

Attali, Y., & Bar-Hillel, M. (2003). Guess where: The position of correct answers in multiple-choice test items as a psychometric variable. *Journal of Educational Measurement*, *40*(2), 109–128. https://doi.org/10.1111/j.1745-3984.2003.tb01099.x

Barnard-Brak, L., Lan, W. Y., & Yang, Z. (2018). Differences in mathematics achievement according to opportunity to learn: A 4pL item response theory examination. *Studies in Educational Evaluation*, *56*, 1–7. https://doi.org/10.1016/j.stueduc.2017.11.002

Crocker, L., & Algina, J. (2008). *Introduction to classical and modern test theory*. Cengage Learning.

Firmansyah, M. A. (2017). Analisis hambatan belajar mahasiswa pada mata kuliah statistika. *Jurnal Penelitian Dan Pembelajaran Matematika*, *10*(2). https://doi.org/10.30870/jppm.v10i2.2036

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Sage.

Hulin, C. L., Drasgow, F., & Parsons, C. K. (1983). *Item response theory: Application to psychological measurement*. Dow Jones-Irwin.

Istiyono, E., Mardapi, D., & Suparno, S. (2014). Pengembangan tes kemampuan berpikir tingkat tinggi fisika (pysthots) peserta didik SMA. *Jurnal Penelitian Dan Evaluasi Pendidikan*, *18*(1), 1–12. https://doi.org/10.21831/pep.v18i1.2120

Kartianom, K., & Mardapi, D. (2018). The utilization of junior high school mathematics national examination data: Conceptual error diagnosis. *REiD (Research and Evaluation in Education)*, *3*(2). https://doi.org/10.21831/reid.v3i2.18120

Kartianom, K., & Ndayizeye, O. (2017). What's wrong with the Asian and African Students' mathematics learning achievement? The multilevel PISA 2015 data analysis for Indonesia, Japan, and Algeria. *Jurnal Riset Pendidikan Matematika*, *4*(2), 200–210.

ISSN 2356-2684 (print), ISSN 2477-1503 (online)

https://doi.org/10.21831/jrpm.v4i2.16931

Keeves, J. P., & Alagumalai, S. (1999). New approaches to measurement. *Advances in Measurement in Educational Research and Assessment*, 23–42.

Kien-Kheng, F., & Idris, N. (2010). A comparative study on statistics competency level using TIMSS data: Are we doing enough? *Journal of Mathematics Education*, *3*(2), 126–138.

Mardapi, D. (2012). *Pengukuran penilaian dan evaluasi pendidikan*. Nuha Medika.

Mills, J. D., & Holloway, C. E. (2013). The development of statistical literacy skills in the eighth grade: Exploring the TIMSS data to evaluate student achievement and teacher characteristics in the United States. *Educational Research and Evaluation*, *19*(4), 323–345. https://doi.org/10.1080/13803611.2013.771110

Muslim, M., Suhandi, A., & Nugraha, M. G. (2017). Development of reasoning test instruments based on TIMSS framework for measuring reasoning ability of senior high school student on the physics concept. *Journal of Physics: Conference Series*, *812*(1), 012108. https://doi.org/10.1088/1742-6596/812/1/012108

Ramos, J. L. S., Dolipas, B. B., & Villamor, B. B. (2013). Higher order thinking skills and academic performance in physics of college students: A regression analysis. *International Journal of Innovative Interdisciplinary Research*, *4*(48–60).

Retnawati, H. (2013). *Evaluasi program pendidikan*. Universitas Terbuka.

Retnawati, H. (2016). *Validitas reliabilitas dan karakteristik butir*. Parama Publishing.

Retnawati, H. (2017). Diagnosing the junior high school students'difficulties in learning mathematics. *International Journal on New Trends in Education and Their Implications*, *8*(1), 33–50. http://www.ijonte.org/FileUpload/ks63207/File/04.heri_retnawati.pdf

Retnawati, H., & Hadi, S. (2014). Sistem bank soal daerah terkalibrasi untuk menyongsong era desentralisasi. *Jurnal Ilmu Pendidikan*, *20*(2), 183–193. https://doi.org/10.17977/jip.v20i2.4615

Rindermann, H., & Baumeister, A. E. E. (2015). Validating the interpretations of PISA and TIMSS tasks: A rating study. *International Journal of Testing*, *15*(1), 276–296. https://doi.org/10.1080/15305058.2014.966911

Rogers, H. J. (1999). Guessing in multiple choice tests. In *Advances in measurement in educational research and assessment* (pp. 235–243). Pergamon Press, New York.

Tee, O. P., & Subramaniam, R. (2018). Comparative study of middle school students' attitudes towards science: Rasch analysis of entire TIMSS 2011 attitudinal data for England, Singapore and the U.S.A. as well as psychometric properties of attitudes scale. *International Journal of Science Education*, *40*(3), 268–290. https://doi.org/10.1080/09500693.2017.1413717

Wibawa, S. (2017). Tri Dharma Perguruan Tinggi (Pendidikan dan pengabdian kepada masyarakat). In *Disampaikan dalam Rapat Perencanaan Pengawasan Proses Bisnis Perguruan Tinggi Negeri. Yogyakarta* (Vol. 29).

Wu, M., Tam, H. P., & Jen, T.-H. (2016). *Educational measurement for applied researchers*. Springer Singapore. https://doi.org/10.1007/978-981-10-3302-5

ISSN 2356-2684 (print), ISSN 2477-1503 (online)