Implementasi *Classic Test* dan *Item Respon Theory* Pada Penilaian Tes Pembelajaran Matematika

Syaiful Syamsuddin
Institut Agama Islam Negeri Curup
e-mail korenpondensi: syaifulsyamsuddin@iaincurup.ac.id

ABSTRACT

The quality of learning assessment is one of the benchmarks in improving the quality of education. A good assessment system will trigger educators to be better at teaching and motivate students to learn better. Assessment of learning outcomes by educators is used to assess the achievement of student competencies and improving the learning process. The form of effort to assess student learning outcomes is to carry out tests by paying attention to the quality of tests based on student abilities. This study uses a quantitative approach to analyze the test items through the Classical Test Theory (CTT) and Item Response Theory (IRT) approaches. This research was conducted by respondents as much as 285 students. The results showed that based on CTT, 22.5% were obtained in the difficult category, 47.5% included in the moderate category, 30% in the easy category and 62.5% of the items already had good discriminating power. As for questions with a distractor function that functions well at 45% with a reliability of 0.785. Meanwhile, with the item IRT, the model that is suitable for this study is IRT 2PL with difficulty levels between -6.337 to 4.945, differential power ranging from 0.271 to 2.254 with 38 fit items.

***KEYWORDS**: Assessment, Measurement, Classical Test Theory, Item Respon Theory*

ABSTRAK

Kualitas penilaian pembelajaran menjadi salah satu tolak ukur dalam peningkatan kualitas pendidikan. Sistem penilaian yang baik akan memicu pendidik untuk lebih baik dalam mengajar dan memotivasi peserta didik untuk belajar yang lebih baik. Penilaian hasil belajar oleh pendidik digunakan untuk menilai pencapaian kompetensi peserta didik dan memperbaiki proses pembelajaran. Bentuk upaya penilaian hasil belajar peserta didik adalah melakukan pengujian dengan memperhatikan kualitas tes berdasarkan kemampuan siswa. Penelitian ini menggunakan pendekatan kuantitatif untuk menganalisis butir tes melalui pendekatan *Classical Test Theory* (CTT) dan *Item Respon Theory* (IRT). Penelitian ini dilakukan dengan melibatkan jumlah responden sebanyak 285 siswa. Hasil penelitian menunjukkan bahwa berdasarkan CTT diperoleh 22,5% dengan kategori sulit, 47,5% termasuk kategori sedang dan sebesar 30% dengan kategori mudah serta 62,5% soal sudah memiliki daya pembeda yang baik. Adapun soal dengan fungsi distraktor yang berfungsi baik sebesar 45% dengan reliabilitas sebesar 0,785. Sedangkan dengan model IRT, model yang sesuai dengan penelitian ini adalah IRT 2PL dengan perolehan tingkat kesukaran antara -6.337 sampai 4.945, daya beda berkisar antara 0,271 sampai 2,254 dengan 38 butir *fit*.

KATA KUNCI: Penilaian, Pengukuran, Teori Tes Klasik, Teori Respon Butir

Article History		
Received: 28 Desember 2022	*Revised: 09 Januari 2023*	*Accepted: 30 Januari 2023*

PENDAHULUAN

Pendahuluan Pada era berkemajuan saat ini, semua negara akan berkompetensi dalam meningkatkan kualitas pendidikan. Melalui pendidikan diharapkan mampu meningkatkan kualitas sumber daya manusia sehingga tingkat kesejahteraan pada masyarakat juga ikut meningkat. Dalam meningkatkan kualitias pendidikan dapat pendidikan dapat ditempuh melalui peningkatan kualitas pembelajaran dan kualitas sistem penilaiannya (Mardapi, 2017). Sejalan dengan hal tersebut, (Utami & Syamsuddin, 2022) menyebutkan bahwa kualitas pembelajaran didasarkan pada standar penilian khususnya pada jenjang pendidikan dasar dan menengah.

Penilaian memiliki peran yang sangat penting dalam pembelajaran. Sebagimana Peraturan Pemerintah Nomor 4 Tahun 2022 tentang Standar Nasional Pendidikan bahwa pelaksanaan pendidikan dikatakan terlaksana apabila sesuai dengan standar nasional pendidikan yang telah ditentukan dan ditetapkan, salah satu diantara 8 standar pendidikan ialah standar penilaian pendidikan (Indonesia, 2022). Standar Penilaian Pendidikan merupakan tolak ukur pada sistem pembelajaran mengenai prinsip, tujuan, manfaat, mekanisme, prosedur serta instrumen yang digunakan untuk menilai hasil belajar siswa (Primasari et al., 2021).

Sistem pembelajaran dan sistem penilian saling terkait. Penilaian merupakan bagian dari suatu proses untuk dapat diketahui seberapa besar tujuan yang akan dicapai (Al-Fraihat et al., 2020). Bila suatu proses pada penilaian tidak sesuai dengan standar yang telah ditetapkan, maka akan terjadi penyederhanaan dalam proses pembelajaran yang diorientasikan dengan bagaimana penilaian itu dilakukan. Secara umum, hasil penilaian merupakan salah satu indikator standar keberhasilan proses pembelajaran pada sistem pendidikan (Kusainun, 2020). (Mukti et al., 2020; Syamsuddin & Istiyono, 2018; Syamsuddin & Setiawati, 2018; Syamsuddin & Utami, 2021; Utami & Syamsuddin, 2020) menyatakan bahwa sistem pembelajaran yang baik akan menghasilkan kualitas belajar yang baik dan kualitas pembelajaran yang baik dapat dilihat dari hasil penilaiannya. Kemudian, sistem penilaian yang baik akan memicu pendidik untuk lebih baik dalam mengajar dan memotivasi peserta didik untuk belajar yang lebih baik. penilaian merupakan kebutuhan instrik dalam kegiatan belajar mengajar (Gronlund, 1968). Untuk itu dalam peningkatan kualitas pembelajaran diperlukan sistem penilaian yang baik pada satuan pendidikan.

Penilaian pada satuan pendidikan secara edukatif merupakan penilaian yang hasilnya digunakan sebagai umpan balik bagi tenaga pendidik, siswa dan orang tua dalam meningkatkan proses pembelajaran dan hasil belajar (Permendikbudristek, 2022). Untuk itu, kegiatan penialain merupaan salah satu kewajiban pendidik untuk melakukannya. Penilaian dilkukan untuk memberikan gambaran sejauhmana kemajuan

peserta didik terhadap kurikululm yang telah diajarkan (Ferreira et al., 2020; González - salamanca et al., 2020). Permendikbudristek Nomor 21 Tahun 2022 bahwa penilaian hasil belajar oleh pendidik dilakukan secara formatif dan sumatif untuk memantau proses, kemajuan, perbaikan hasil, mengumpulkan informasi mengenai kesulitan belajar serta perkembangan peseta didik. Sejalan dengan hasil penelitian (Ikhwan, 2013) bahwa penilaian hasil belajar oleh pendidik digunakan untuk menilai pencapaian kompetensi peserta didik; bahan penyusunan laporan hasil belajar; dan memperbaiki proses pembelajaran. Berdasarkan hal tersebut dapat diartikan bahwa salah satu upaya penilaian hasil belajar peserta didik adalah dengan memberikan tes hasil belajar. Meskipun, terkadang tes yang diberikan kepada peserta didik belum mampu menunjukkan kemampuan peserta didik itu sendiri (Andayani et al., 2019; Purnama & Alfarisa, 2020). Oleh karena itu, diperlukan suatu metode dalam menganalisis butir tes hasil belajar peserta didik yang mampu memnunjukkan kemauan peserta didik.

Terdapat dua metode yang dapat digunakan untuk menganalisis butir tes yaitu dengan pendekatan Teori Tes Klasik (*Classical Test Theory*) dan pendekatan Teori Respon Butir (*Item Respon Theory*) (Abdu Bichi et al., 2015; Jabrayilov et al., 2016). Teori Tes Klasik (CTT) dan Teori Respon Butir (IRT) umumnya dianggap sebagai dua kerangka statistik yang digunakan untuk mengatasi hal yang berkaitan dengan pengukuran. Baik CTT maupun IRT menggambarkan karakteristik individu, menganalisis kemampuan dan atribut laten serta memungkinkan untuk memprediksi hasil psikologis dan tes pendidikan dengan mengidentifikasi parameter item dalam hal ini tingkat kesukaran, daya beda dan kemampuan peserta tes (Abdu Bichi et al., 2015).

(Hambleton & Jones, 1993) menjelaskan bahwa CTT mudah diterapkan dalam banyak situasi pengujian, dimana kemampuan seseorang bergantung pada item dan statistik item yang meliputi tingkat kesukaran dan daya beda bergantung pada sampel. (Cappelleri et al., 2014) berpendapat bahwa CTT merupakan pendekatan yang mudah dipahami dan sederhana dalam menganalisis tes secara empirik yang jika digambarkan, kemampuan peserta tes dilaporkan dalam hal jumlah butir yang dijawab benar.

Di sisi lain, IRT lebih berlandaskan teori dan memodelkan distribusi keberhasilan peserta tes di tingkat butir. IRT berfokus pada informasi tingkat item berbeda dengan fokus utama CTT pada informasi tingkat tes (R.K Hambleton & Swaminathan, 1985) menjelaskan bahwa Item Respon Theory (IRT) merupakan salah satu cara untuk menilai kelayakan butir dengan membandingkan rerata penampilan butir terhadap tampilan bukti kemampuan kelompok yang diramalkan oleh model. Secara sederhana dijelaskan (Hambleton & Jones, 1993) bahwa IRT sebagai teori statistik umum tentang item yang diuji dan performa tes dan bagaimana performa berhubungan dengan kemampuan yang diukur oleh item dalam tes.

Ada beberapa kriteria yang diperhatikan dalam CTT yaitu tingakat kesukaran, daya beda, efektivitas distractor dan reliabilitas skor tes (Hamimi et al., 2020; Suwarto,

2021). Akan tetapi, CTT dianggap kurang maksimal dalam menggambarkan kemampuan peserta tes yang sebenarnya (Amelia & Kriswantoro, 2017). Oleh karena itu, IRT hadir bertujuan untuk mengatasi kelemahan pengukuran melalui CTT, yang berarti sebuah tes dinilai berdasarkan masing-masing item. Sehingga setiap butir memiliki tingkat kesulitamn yang berbeda, memperhitungkan kemampuan peserta didik, serta karakteristik tes tidak bergantung terhadap peserta tes (Purnama & Alfarisa, 2020).

Pada teori respon butir digunakan model matematis dalam menghubungkan karakteristik butir soal dengan kemampuan responden. (Retnawati, 2014) menyatakan bahwa model matematis pada teori respon butir memiliki makna bahwa probabilitas subjek untuk menjawab butir dengan benar tergantung pada kemampuan subjek dan karakteristik butir. Tiga model IRT yang paling umum digunakan adalah model logistik satu parameter (model 1PL atau Rasch), model logistik dua parameter (2PL) dan model logistik tiga parameter (3PL). Ketiga model tersebut memperhatikan tingkat kesukaran butir. Selain itu, model 2PL dan 3PL memngukur daya beda yang memungkinkan butir untuk membedakan kemampuan peserta tes. Sedangkan Model 3PL berisi disebut sebagai parameter untuk melihat guessing yang terjadi pada karakteristik butir (Abdu Bichi et al., 2015).

METODE

Metode Pada Penelitian ini merupakan penelitian dekriptif dengan pendekatan kuantitatif yang bertujuan untuk memberikan gambaran mengenai hasil analisis butir tes melalui pendekatan Teori Tes Klasik (*Classic Theory*) dan Teori Respon Butir (*Item Respon Theory*). Penelitian ini melibatkan 285 siswa SMA se-derajat di Yogyakarta. Siswa yang terlibat dalam penelitian ini ditentukan melalui teknik *purposive sampling* yang selanjutnya mengerjakan 40 butir tes pembelajaran matematika.

Tahapan atau prosedur yang dilakukan dalam penelitian ini adalah sebagai berikut: (1) Penyiapan data berupa penginputan jawaban peserta tes; (2) Penerapan teori uji klasik dengan menghitung indeks tingkat kesukaran, daya beda, sebaran pilihan jawaban dan reliabilitas soal; (3) Penerapan teori respon butir dengan menggunakan model IRT 1PL, 2PL dan 3PL dengan pengujian asumsi model IRT, menghitung parameter karakteristik butir soal(tingkat kesukaran dan daya beda), menghitung parameter kemampuan untuk setiap model, mencari model ICC yang sesuai dari setiap soal dalam setiap model, mencari model yang paling sesuai untuk menggambarkan setiap soal

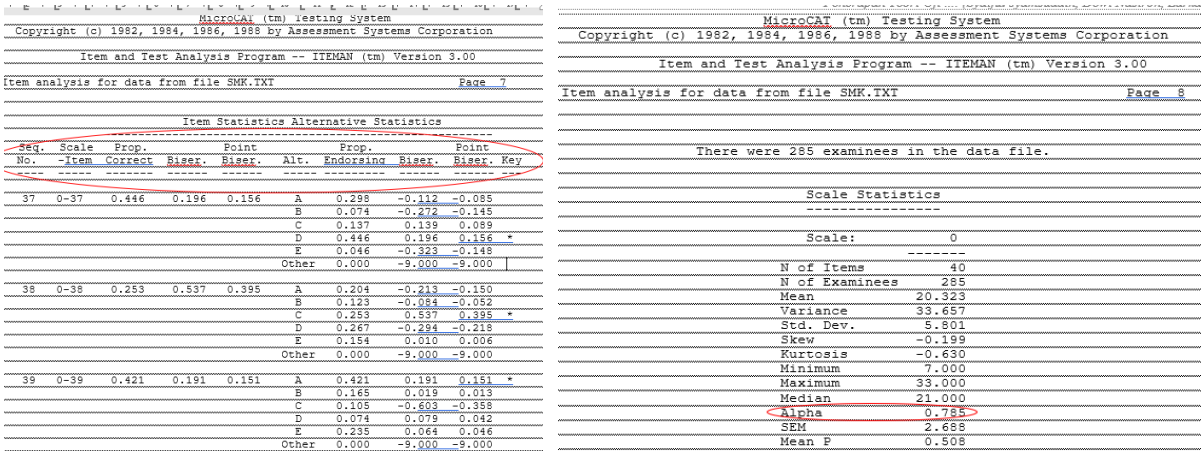
Software yang digunakan pada penelitian ini adalah ITEMAN dan BILOG-MG. Program ITEMAN digunakan untuk analisis teori tes klasik (classic theory), sedangkan program BILOG-MG digunakan untuk menganalisis model item respon theory (IRT)

1PL, 2PL dan 3PL

HASIL dan PEMBAHASAN

Classic Test Theory (CTT)

Beberapa aspek yang diperhatikan dalam teori uji klasik yaitu tingkat kesukaran, daya beda, kebermanfaatan distraktor dan reliabilitas skor tes (Perdana, 2018). Hasil analisis butir CTT melalui program ITEMAN memberikan gambaran karakteristik butir meliputi tingkat kesukaran, daya beda dan keberfungsian distractor serta reliabilitas skor tes, sebagaimana gambar yang disajikan berikut ini:



MicroCAT (tm) Testing System
Copyright (c) 1982, 1984, 1986, 1988 by Assessment Systems Corporation
Item and Test Analysis Program -- ITEMAN (tm) Version 3.00
Item analysis for data from file SMK.TXT Page 7

Item Statistics Alternative Statistics

Seq. No.	Scale -Item	Prop. Correct	Biser.	Point Biser.	Alt.	Prop. Endorsing	Biser.	Point Biser.	Key
37	0-37	0.446	0.196	0.156	A	0.298	-0.112	-0.085	
					B	0.074	-0.272	-0.145	
					C	0.137	0.139	0.089	
					D	0.446	0.196	0.156	*
					E	0.046	-0.323	-0.148	
					Other	0.000	-9.000	-9.000	
38	0-38	0.253	0.537	0.395	A	0.204	-0.213	-0.150	
					B	0.123	-0.084	-0.052	
					C	0.253	0.537	0.395	*
					D	0.267	-0.294	-0.218	
					E	0.154	0.010	0.006	
					Other	0.000	-9.000	-9.000	
39	0-39	0.421	0.191	0.151	A	0.421	0.191	0.151	*
					B	0.165	0.019	0.013	
					C	0.105	-0.603	-0.358	
					D	0.074	0.079	0.042	
					E	0.235	0.064	0.046	
					Other	0.000	-9.000	-9.000	

MicroCAT (tm) Testing System
Copyright (c) 1982, 1984, 1986, 1988 by Assessment Systems Corporation
Item and Test Analysis Program -- ITEMAN (tm) Version 3.00
Item analysis for data from file SMK.TXT Page 8

There were 285 examinees in the data file.

Scale Statistics

Scale:	0
N of Items	40
N of Examinees	285
Mean	20.323
Variance	33.657
Std. Dev.	5.801
Skew	-0.199
Kurtosis	-0.630
Minimum	7.000
Maximum	33.000
Median	21.000
Alpha	0.785
SEM	2.688
Mean P	0.508

Gambar 1. *Output* Analisis Butir melalui CTT menggunakan ITEMAN

Tingkat kesukaran butir soal memberikan gambaran mengenai kemungkinan seberapa banyak responden menjawab butir tes dengan benar (Erfan et al., 2020). Adapun kriteria tingkat kesukaran butir dengan pendekatan CTT disajikan pada kriteria sebagai berikut (Lestari & Yudhanegara, 2017):

Tabel 1. Kriteria Indeks Kesukaran

Kategori	Interpretasi Indeks Kesukaran
Sukar	$0.00 < IK \leq 0.30$
Sedang	$0.30 < IK \leq 0.70$
Mudah	$0.70 < IK \leq 1.00$

Tingkat kesukaran butir tes ini dapat dilihat melalui output ITEMAN pada kolom prop correct (Allen & Yen, 1979). Berdasarkan kriteria yang tersaji pada tabel 1 di atas diperoleh 22,5% butir dengan kategori sukar, 47,5% termasuk kategori sedang dan sebesar 30% dengan kategori mudah. Hasil analisis rangkuman butir ditinjau dari tingkat kesukarannya dilihat pada tabel 2 berikut ini:

Tabel 2. Kriteria Indeks Kesukaran

Kategori	No. butir
Sukar	10, 22, 23, 24, 26, 27, 29, 30, 31
Sedang	2,4,5, 9, 17, 18, 19, 21, 25, 28, 32, 33, 34, 35, 36, 37, 38,39, 40
Mudah	1,3,6,7,8, 11, 12, 13, 14, 15,16, 20

Persebaran tingkat kesukaran menggunakan CTT di atas telah merata. Sebanyak 40 butir yang dianalisis, tidak seluruh butir memiliki karakteristik yang sukar ataupun mudah, melainkan tersebar dengan baik. (Arikunto, 2012) menjelaskan bahwa tes yang baik adalah tes yang memiliki persebaran butir yang tidak terlalu mudah ataupun tidak terlalu sulit. Butir terlalu mudah tidak dapat memicu peserta didik untuk untuk meningkatkan usaha dalam pemecahannya. Sebaliknya butirl yang terlalu sukar menimbulkan rasa putus asa dan menurunkan motivasi peserta didiki untuk mengulangi tes karena diluar batas kemampuannya (Suzana, 2018).

Kriteria selanjutnya yang perlu diperhatikan dalam analisis butir menggunakan CTT yakni indeks daya beda. Daya beda merupakan kemampuan tes untuk membedakan peserta tes yang memiliki kemampuan tinggi dan kemampuan rendah berdasarkan kemampuan peserta tes menjawab soal (Suwarto, 2022).

Daya beda butir biasanya dilakukan dengan menggunakan indeks korelasi, diskriminasi, dan indeks keselarasan item. Dari ketiga cara tersebut yang paling sering digunakan adalah indeks korelasi (Yen, 1992). Hasil analisis daya beda butir soal dilihat pada output point biserial sebagaimana penelitian sebelumnya (Saputra et al., 2021). Adapun kriteria suatu butir dikatakan baik ketika indeks daya beda lebih dari 0.20 dengan kriteria cukup (Erfan et al., 2020; Suwarto, 2022). Berikut rangkuman hasil analis indeks daya beda butir soal yang disajikan pada tabel di bawah ini:

Tabel 3. Daya Beda Berdasarkan CTT

Kategori	No Butir
Baik	2,3, 5, 6, 9,11, 12,13, 14, 15, 16,17,18,19,22,23, 24, 25, 31, 32, 33, 34, 35, 36, 38
Belum baik	1,4, 7, 8, 10,20, 21,26, 27,28, 29, 30, 37, 39, 40

Dari pembacaan tabel 2 diperoleh bahwa 62,5% soal sudah memiliki daya pembeda yang baik dan selebihnya sebanyak 37,5% belum mampu memiliki daya pembeda yang baik sehingga masih diperlukan adanya perbaikan. Hasil ini menunjukkan bahwa 15 butir dari 40 butir yang diujikan harus direvisi. Sebagaimana (Dichoso & Joy, 2020) menyebutkan bahwa butir tes yang berada dalam kategori belum baik harus di revisi dan ketika nilai indeks daya beda sangat rendah maka harus di keluarkan. Meskipun

demikian, butir-butir yang berada dalam kategori baik dapat digunakan unutk tes pembelajaran matematika. Hal tersebut juga memberikan informasi jika butir-butir tes pembelajaran matematika ini dapat membedakan siswa yang memiliki kemampuan tinggi, sedang ataupun rendah. Semakin besar indeks daya beda butir tes maka butir tersebut mampu mendeteksi perbedaan individu diantara siswa (Singh et al., 2014; Suwarto, 2021).

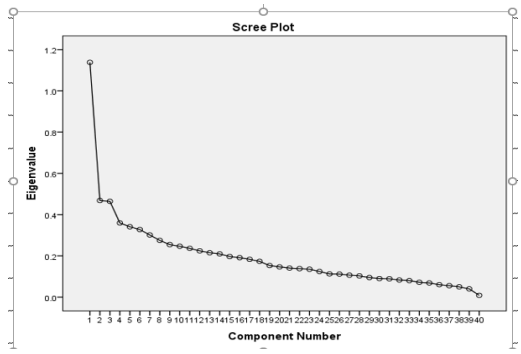
Distraktor atau biasa disebut sebagai pengecoh yang berarti jawaban yang bernilai salah dari bentuk tes pilihan ganda. Distraktor memiliki fungsi sebagai opsi pengecoh yang dapat membuat peserta tes merasa kebingungan dalam memilih jawaban benar diantara alternative jawaban yang disediakan (Suwarto, 2022). Suatu distraktor dapat dikatakan berfungsi dengan baik jika dipilih paling sedikit 5% untuk minimal 4 pilihan jawaban (Kementerian Pendidikan Dan Kebudayaan, 2017; Suwarto, 2021). Berdasarkan hasil analis dengan ITEMAN diperoleh informasi bahwa 50% butir memiki distraktir yang berfungsi dengan baik meliputi butir 7, 8,9,10,14,19,22, 23, 24, 26, 27, 28, 29, 30, 33, 34, 38,39, 40. Sedangkan 50% butir lainnya, keberfungsian distraktor tidak berjalan dengan baik atau dengan kata lain 50% pengecoh butir-butir tersebut harus direvisi. Sebagaimana penelitian (Maharani & Putro, 2020) bahwa 20% butir dengan distraktor yang tidak efektif harus direvisi.

Reliabilitas skor tes dilihat menggunakan koefisien alpha. Nilai koefisien alfa yang diperoleh dari hasil ITEMAN sebesar 0,885. Hal ini menunjukan bahwa butir-butir tes te;ah memenuhi kriteria reliabel. Suatu tes dapat dikatakan reliabel jika nilai koefisien alpha berada di atas batas nilai koefisien reliabilitas 0.70 (Pascual & North, 2016; Sugianto, 2017).

Analisis butir menggunakan *Classic Test Theory* (CTT) memiliki kekurangan dikarenakan butir tes bergantung pada responden yang dikenai butir tes (Amelia & Kriswantoro, 2017). Lebih lanjut dijelaskan bahwa pada CTT, tes terasa mudah jika dikerjakan oleh responden dengan kemampuan tinggi (indeks kesukaran butir menjadi besar), sedangkan responden dengan kemampuan rendah maka tes akan terasa sukar (indeks kesukaran butir menjadi kecil) (Suwarto, 2022). Dengan demikian, peneliti imelanjutkan analisis karakteristik butir soal dengan pendekatan *Item Respon Theory* (IRT).

***Item Respon Theory* (IRT)**

Unidimensional merupakan salah satu prasyarat yang harus dipenuhi dalam model IRT Hambleton & Swaminathan, 1985:16). Retnowati (2014) menerangkan bahwa salah satu cara yang digunkan untuk mengetahui asumsi unidimensional adalah dengan analisis faktor. Berikut gambar hasil analisis faktor yang dipereoleh dengan mengguakan SPSS disajikan melalui gambar berikut:



Gambar 2. Grafik Unidimensional Tes Pembelajaran Matematika

Scree plot diatas memberikan informasi bahwa faktor yang terbentuk adalah 1 faktor yang berarti unidimensio. (Saepuzaman et al., 2021) menaksir asumsi *unidimensional* berdasarkan pada rasio akar ciri pertama dan kedua. Jika perbandingan nilai perbandingan akar ciri yang tinggi mengindikasikan unidimensional. Gambar diatas menunjukkan penurunan grafik yang ekstrim antara faktor dan faktor 2 hingga hampir membentuk sudut siku-siku. Hal ini berarti hanya terdapat 1 faktor dominan dalam perangkat tes pembelajaran matematika atau dengan kata lain diartikan jika setiap butir hanya mengukur satu kemampuan (Retnawati, 2014). Faktor pertama memiliki akar ciri (*eigenvalue*) sebesar 1,138 sedangkan faktor-faktor lainnya memiliki akar ciri (*eigenvalue*) kurang dari satu.

Terdapat dua cara yang bisa digunakan untuk menentukan kecocokan model parameter logistik yakni dengan metode grafik dengan memperhatikan nilai *chi square* (x^2) dan metode grafik dengan melihat kurva ICC (Retnawati, 2014). Lebih lanjut dijelaskan, (R.K Hambleton & Swaminathan, 1985; Saepuzaman et al., 2021) bahwa butir dikatakan cocok jika nilai *chi square* (x^2) atau *Threshold* pada *output* ITEMAN berada pada rentang -2 s.d +2 (Cheng et al., 2019).

Pada penelitian ini, kecocokan model menggunakan metode statistik dan metode grafik. Adapun dalam menentukan model parameter yang cocok untuk kedua metode (statistik dan grafik) yakni dengan membandingkan paling banyak butir yang cocok (*fit*) dengan model parameter logistik (1 PL, 2PL dan 3PL) (Saepuzaman et al., 2021). Berikut rangkuman kecocokan butir 1PL, 2PL dan 3 PL ditinjau dari tingkat kesukaran disajikan pada tabel 4 berikut ini:

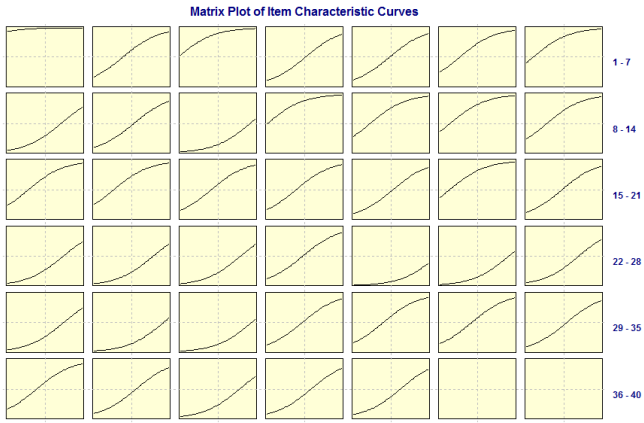
Tabel 4. Kecocokan Butir Pada Model Parameter (1PL, 2PL, dan 3PL)

No. Butir	1 PL	2 PL	3PL

	Threshold	Ket	*Threshold*	Ket	*Threshold*	Ket
1	-6.804	Tidak Fit	-6.337	Tidak Fit	-5.76	Tidak Fit
2	-0.706	Fit	-0.802	Fit	-0.071	Fit
3	-3.205	Tidak Fit	-1.901	Fit	-1.634	Fit
4	0.061	Fit	0.075	Fit	1.16	Fit
5	-0.117	Fit	-0.121	Fit	0.452	Fit
6	-1.425	Fit	-0.93	Fit	-0.52	Fit
7	-2.455	Tidak Fit	-2.328	Tidak Fit	-1.725	Fit
8	1.249	Fit	1.467	Fit	1.773	Fit
9	0.172	Fit	0.182	Fit	0.835	Fit
10	2.528	Tidak Fit		-		Tidak Fit
11	-2.957	Tidak Fit	-2.041	Fit	-1.657	Fit
12	-1.674	Fit	-1.197	Fit	-0.842	Fit
13	-2.204	Tidak Fit	-1.502	Fit	-1.098	Fit
14	-1.372	Fit	-1.156	Fit	-0.574	Fit
15	-1.365	Fit	-0.692	Fit	-0.515	Fit
16	-1.452	Fit	-0.936	Fit	-0.423	Fit
17	-0.568	Fit	-0.371	Fit	-0.064	Fit
18	-0.706	Fit	-0.505	Fit	-0.184	Fit
19	0.128	Fit	0.137	Tidak Fit	0.788	Fit
20	-2.239	Tidak Fit	-2.231	Fit	-1.65	Fit
21	-0.161	Fit	-0.287	Fit	1.172	Fit
22	1.431	Fit	0.977	Fit	1.103	Fit
23	1.706	Fit	1.607	Fit	1.665	Fit
24	1.764	Fit	1.438	Fit	1.524	Fit
25	-0.161	Fit	-0.087	Fit	0.134	Fit
26	3.669	Tidak Fit	3.456	Tidak Fit	2.789	Tidak Fit
27	2.49	Tidak Fit	4.945	Tidak Fit	-	Tidak Fit
28	1.173	Fit	2.55	Tidak Fit	-	Tidak Fit
29	1.458	Fit	3.467	Tidak Fit	-	Tidak Fit
30	2.528	Tidak Fit	4.838	Tidak Fit	-	Tidak Fit
31	2.646	Tidak Fit	1.952	Fit	-	Tidak Fit
32	-0.206	Fit	-0.168	Fit	-	Tidak Fit
33	-0.637	Fit	-0.422	Fit	-	Tidak Fit
34	-0.476	Fit	-0.539	Fit	-	Tidak Fit
35	0.217	Fit	0.153	Fit	-	Tidak Fit
36	-0.683	Fit	-0.539	Fit	-	Tidak Fit
37	0.351	Fit	0.611	Fit	-	Tidak Fit
38	1.706	Fit	1.478	Fit	-	Tidak Fit
39	0.509	Fit	1.009	Fit	-	Tidak Fit
40	0.739	Fit	1.025	Fit	-	Tidak Fit

Berdsarkan hasil analisis di atas karakteristik keseluruhan butir tes pembelajaran matematika menggunakan model IRT 1PL menunjukan bahwa soal tersebut memiliki

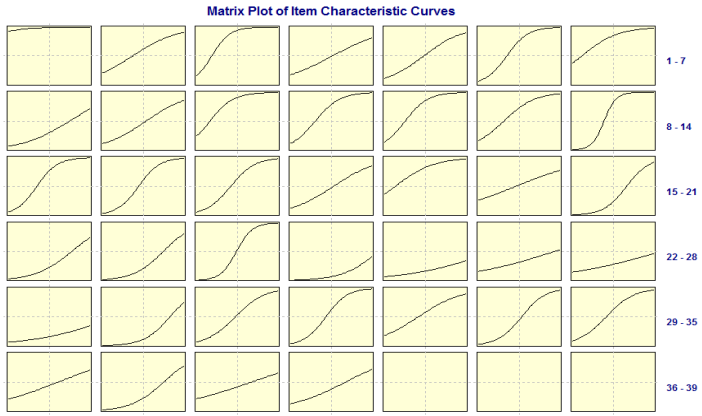
rentang tingkat kesukaran antara -6.804 sampai 3.669. Nilai *chi-square* atau *threshold* butir yang berada dalam rentang *fit* berkisar 72.5% meliputi butir 2, 4, 5, 6, 8, 9 12, 14, 15, 16, 17, 18, 19, 21, 22, 23, 24, 25, 28, 29, 32, 33, 34, 35, 36, 37, 38, 39, dan 40. Selain itu, model kurva ICC (metode grafik) untuk model *parameter logistic* 1 PL dapat dilihat pada gambar di bawah ini:



Gambar 3. Kurva ICC 1 PL Tes Pembelajaran Matematika

Hasil analisa karakteristik butir tes pembelajaran matematika menggunakan model IRT 2PL menunjukan bahwa soal tersebut mempunyai tingkat kesukaran antara -6.337 sampai 4.945. atau dengan kata lain 77.5% keseluruhan butir cocok dengan model meliputi butir 2, 3, 4, 5, 6, 8, 9, 11, 12, 13,14, 15, 16, 17, 18, 19 20, 21, 22, 23, 24, 25, 31, 32, 33, 34, 35, 36, 37, 38, 39, dan 40. Adapun daya beda pada butir tes tersebut berkisar antara 0,271 sampai 2,254. (R.K Hambleton & Swaminathan, 1985) menyebutkan jika daya beda butir soal yang baik berada pada kisaran 0 s.d 3.

Berdasarkan karakteristik tersebut, keseluruhan butir memiliki daya beda yang baik. Hasil ini sejalan dengan penelitian lain (Danuwijaya, 2018; Saputra et al., 2021) bahwa dalam hasil penelitiannya memperoleh indeks daya beda yang baik meskipun konteks penelitiannya berbeda. Butir-butir yang memiliki indeks daya beda yang baik dapat membedakan responden atau peserta tes yang berkemampuan tinggi dan berkemampuan rendah (Uddin et al., 2020) dalam konteks tes pembelajaran mataematika. Selain itu, model kurva ICC (metode grafik) untuk model parameter logistic 2 PL dapat dilihat pada gambar di bawah ini:



Gambar 4. Kurva ICC 2 PL Tes Pembelajaran Matematika

Analisis berikutnya yang dilakukan yaitu mengestimasi parameter butir dengan model 3PL meliputi tingkat kesukaran, daya beda dan *guessing*. Hasil analisa karakteristik butir tes pembelajaran matematika menggunakan model IRT 3PL menunjukan bahwa analisis butir soal tidak bisa dilakukan sampai 40 butir, hanya 25 butir yang dianalisis dari sejumlah butir yang ada yakni 40 butir soal. Hal ini terjadi karena batas acuan untuk memberhentikan penyajian soal pada pada *output* ITEMAN kesalahan baku pengukuran (SE) adalah 0.01 (Fatkhudin et al., 2016).

Model IRT dengan 3 PL memperhatikan nilai *asimtot* bawah pada *output* ITEMAN (Allen & Yen, 1979; R.K Hambleton & Swaminathan, 1985) tidak sama dengan 0 yang mengindikasikan adanya unsur *guessing* (tebakan) (Suwarto, 2011). Mengaju hal tersebut, pada model 3 PL diperoleh indeks tingkat kesukaran butir berada pada rentang -5.760 sampai 2.789, daya beda berkisar antara 0.497 sampai 2.595 dan 26 butir yang teranalisis dalam penelitian ini tidak sama dengan 0. Berdasarkan hal tersebut maka 57.5% butir *fit* atau yang cocok dengan model meliputi butir 2, 3, 4, 5, 6, 7, 8, 9, 11, 12, 13, 14, 15, 16, 17, 18, 20, 21, 22, 23, 24, 25, 26.

KESIMPULAN dan SARAN

Kesimpulan Berdasarkan hasil pembahasan terdapat perbedaan parameter daya pembeda dan tingkat kesukaran pada metode *Classic Test Theory* dan Item *Respon Theory* (IRT). Pada kasus CTT dipengaruhi oleh kemampuan kelompok.yang berarti kemampuan peserta tes/responden dinyatakan pada variabel yang bersifat diskrit serta nilai koefisien reliabilitas tergantung pada peserta tes yang mengikuti tes. Sedangkan dalam kasus IRT dipengaruhi oleh kemampuan individu yang berarti kemampuan peserta tes/responden dinyatakan pada variabel yang sifatnya kontinu serta dalam menentukan koefisien reliabilitas tidak memerlukan tes paralel serta tidak bergantung kepada responden/peserta tes yang mengikuti tes. Meskipun demikian, dalam pengujian

menggunakan CTT tidak mewajibkan jumlah sampel yang besar dan lebih mudah dipahami. Sebaliknya, analisis butir soal dengan IRT memerlukan jumlah sampel yang besar untuk hasil analisis yang lebih *representative* dan memerlukan *software* yang baik untuk melakukan estimasi parameter yang akurat.

Model yang paling sesuai untuk menggambarkan 40 butir tes pada pembelajaran matematika adalah model teori respon butir 2PL. Item fit pada model teori respon butir 2PL antara lain 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40 yaitu berkisar 77.5% butir fit.

Saran yang dapat diberikan penelitian yang dilakukan memperhatikan jumlah responden. Analisis dengan IRT harus memiliki jumlah responden yang lebih banyak dan dapat dianalisis dengan menggunakan analisis faktor dengan memperhatikan KMO dan *Bartlett's test*. Selain itu, tingkat kesukaran butir sebaiknya dibuat proportional dengan peresentase 25% kategori mudah, 25% sulit dan 50% lainnya memiliki kategori mudah

DAFTAR RUJUKAN

Abdu Bichi, A., Embong, R., & Mamat, M. (2015). Comparison of Classical Test Theory and Item Response Theory:A Review of Empirical Studies. *Australian Journal of Basic and Applied Sciences*, *9*(7), 549–556.

Al-Fraihat, D., Joy, M., Masa'deh, R., & Sinclair, J. (2020). Evaluating E-learning systems success: An empirical study. *Computers in Human Behavior*, *102*, 67–86. https://doi.org/10.1016/j.chb.2019.08.004

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

Amelia, R. N., & Kriswantoro, K. (2017). Implementation of Item Response Theory for Analysis of Test Items Quality and Students' Ability in Chemistry. *JKPK (Jurnal Kimia Dan Pendidikan Kimia)*, *2*(1), 1. https://doi.org/10.20961/jkpk.v2i1.8512

Andayani, A., Purwanto, & Ramalis, T. R. (2019). Kajian implementasi teori respon butir dalam menganalisis instrumen tes materi fisika. *Prosiding Seminar Nasional Fisika 5.0*, *1*(1), 37–42.

Arikunto, S. (2012). *Dasar - Dasar Evaluasi Pendidikan*. Bumi Aksara.

Cappelleri, J. C., Jason Lundy, J., & Hays, R. D. (2014). Overview of classical test theory and item response theory for the quantitative assessment of items in developing patient-reported outcomes measures. *Clinical Therapeutics*, *36*(5), 648–662. https://doi.org/10.1016/j.clinthera.2014.04.006

Cheng, Y., Yang, Y., & Du, D. Z. (2019). A class of asymptotically optimal group testing

strategies to identify good items. *Discrete Applied Mathematics*, *260*, 109–116. https://doi.org/10.1016/j.dam.2019.01.040

Danuwijaya, A. A. (2018). Item Analysis of Reading Comprehension Test for Post-Graduate Students. *English Review: Journal of English Education*, *7*(1), 29. https://doi.org/10.25134/erjee.v7i1.1493

Dichoso, A. A., & Joy, M. R. J. (2020). Test item analyzer using point-biserial correlation and p-values. *International Journal Of Scientific & Technology Research*, *9*(4), 2122–2126.

Erfan, M., Maulyda, M. A., Hidayati, V. R., Astria, F. P., & Ratu, T. (2020). Analisis Kualitas Soal Kemampuan Membedakan Rangkaian Seri dan Paralel melalui Teori Tes Klasik Dan Model Rasch. *Indonesian Journal of Educational Research and Review*, *3*(1), 11–19.

Fatkhudin, A., Surarso, B., & Subagio, A. (2016). Item Response Theory Model Empat Parameter Logistik Pada Computerized Adaptive Test. *Jurnal Sistem Informasi Bisnis*, *4*(2), 121–129. https://doi.org/10.21456/vol4iss2pp121-129

Ferreira, M., Martinsone, B., & Talić, S. (2020). Promoting Sustainable Social Emotional Learning at School through Relationship-Centered Learning Environment, Teaching Methods and Formative Assessment. *Journal of Teacher Education for Sustainability*, *22*(1), 21–36. https://doi.org/10.2478/jtes-2020-0003

González-salamanca, J. C., Agudelo, O. L., & Salinas, J. (2020). Key competences, education for sustainable development and strategies for the development of 21st century skills. A systematic literature review. *Sustainability (Switzerland)*, *12*(24), 1–17. https://doi.org/10.3390/su122410366

Gronlund, N. . (1968). *Constructing Achievement Tests Third Edition*. New Jersey:Prentice-Hall.

Hambleton, R.K, & Swaminathan, H. (1985). *Item response theory: principlesand applications*. MA: Kluwer-Nijhoff.

Hambleton, Ronald K, & Jones, R. W. (1993). Comparison of classical test theory and item response theory and. *Educational Measurement*, *12*(3), 38–47. papers2://publication/uuid/A3D74B30-9CF1-4A78-83BE-D6650B671ED1

Hamimi, L., Zamharirah, R., & Rusydy, R. (2020). Analisis Butir Soal Ujian Matematika Kelas VII Semester Ganjil Tahun Pelajaran 2017/2018. *Mathema: Jurnal Pendidikan Matematika*, *2*(1), 57. https://doi.org/10.33365/jm.v2i1.459

Ikhwan, A. (2013). *Management of Learning Assessment Using Curriculum 2013 (Case Study in Islamic Primary School (MI) Muhammadiyah 5 Wonosari Ponorogo - East Java - Indonesia*. *08*(02), 108–123.

Indonesia, P. (2022). Peraturan Pemerintah Republik Indonesia Nomor 4 Tahun 2022 Tentang Perubahan Atas Peraturan Pemerintah Nomor 57 Tahun 2021 Tentang Standar Nasional Pendidikan. *Lembaran Negara Republik Indonesia Nomor 14 Tahun 2022*, 1–16. https://peraturan.bpk.go.id/Home/Details/196151/pp-no-4-tahun-2022

Jabrayilov, R., Emons, W. H. M., & Sijtsma, K. (2016). Comparison of Classical Test Theory

and Item Response Theory in Individual Change Assessment. *Applied Psychological Measurement*, *40*(8), 559–572. https://doi.org/10.1177/0146621616664046

Kementerian Pendidikan Dan Kebudayaan. (2017). Panduan Penilaian oleh Pendidik dan Satuan Pendidikan Sekolah Menengah Pertama. *Kementerian Pendidikan Dan Kebudayaan Direktorat Jenderal Pendidikan Dasar Dan Menengah*, 43–45. http://repositori.kemdikbud.go.id/18051/1/1. Panduan Penilaian SMP - Cetakan Keempat 2017.pdf

Kusainun, N. (2020). Analisis Standar Penilaian Pendidikan di Indonesia. *Jurnal Keislaman Dan Kemasyarakatan*, *4*(1), 134–154.

Lestari, K. E., & Yudhanegara, M. R. (2017). *Penelitian Pendidikan Matematikan (Anna (ed.)*. PT. Refika Aditama.

Maharani, A. V., & Putro, N. H. P. S. (2020). Item Analysis of English Final Semester Test. *Indonesian Journal of EFL and Linguistics*, *5*(2), 491. https://doi.org/10.21462/ijefl.v5i2.302

Mardapi, D. (2017). Pengukuran, Penilaian, dan Evaluasi Pendidikan. In *Academia Edu* (Vol. 7, Issue 2). Yogyakarta Nuha Medika.

Mukti, T. S., Utami, M. A. P., & Puspitasari, F. F. (2020). Sekolah Alam: Evaluasi Program Sekolah dalam Menumbuhkan Kecerdasan Naturalistik dan Kinestetik pada Pendidikan Anak Usia Dini. *INSANIA : Jurnal Pemikiran Alternatif Kependidikan*, *25*(1), 123–132. https://doi.org/10.24090/INSANIA.V25I1.3542

Pascual, G. R., & North, C. (2016). Analysis of The English Achievement Test for ESL Learners in Northern Philippines. *International Journal of Advanced Research in Management and Social Sciences*, *5*(12), 1–5. www.garph.co.uk

Perdana, S. A. (2018). Analisis Kualitas Instrumen Pengukuran Pemahaman Konsep Persamaan Kuadrat Melalui Teori Tes Klasik Dan Rasch Model. *Jurnal Kiprah*, *6*(1), 41–48. https://doi.org/10.31629/kiprah.v6i1.574

Permendikbudristek. (2022). Standar Penilaian Pendidikan Permendikbudristek No 21 tahun 2022. *Gurusumedang.Com*. https://www.gurusumedang.com/2022/06/standar-penilaian-pendidikan.html

Primasari, I. F. N. D., Marini, Arita, S., & Mohamad, S. (2021). Analisis Kebijakan Dan Pengelolaan Pendidikan Terkait Standar Penilaian Di Sekolah Dasar. *Jurnal Basicedu*, *5*(3), 1479–1491. https://jbasic.org/index.php/basicedu/article/view/956

Purnama, D. N., & Alfarisa, F. (2020). Karakteristik Butir Soal Try Out Teori Kejuruan Akuntansi Smk Berdasarkan Teori Tes Klasik Dan Teori Respons Butir. *Jurnal Pendidikan Akuntansi Indonesia*, *18*(1), 36–46. https://doi.org/10.21831/jpai.v18i1.31457

Retnawati, H. (2014). *Teori Respon Butir dan Penerapannya*. Nuha Medika.

Saepuzaman, D., Istiyono, E., Haryanto, H., Retnawati, H., & Yustiandi, Y. (2021). Analisis

Karakteristik Butir Soal Fisika Dengan Pendekatan IRT Penskoran Dikotomus dan Politomus. *Radiasi : Jurnal Berkala Pendidikan Fisika*, *14*(2), 62–75. https://doi.org/10.37729/radiasi.v14i2.1200

Saputra, A. N. S., Retnawati, H., & Yusron, E. (2021). Analysis Difficulties and Characteristics of Item Test of on Biology National Standard School Examination. *Proceedings of the 6th International Seminar on Science Education (ISSE 2020)*, *541*(Isse 2020), 8–14. https://doi.org/10.2991/assehr.k.210326.002

Singh, J. P., Kariwal, P., Gupta, S. B., & Shrotriya, V. P. (2014). Improving Multiple Choice Questions (MCQs) through item analysis : An assessment of the assessment tool. *International Journal of Sciences & Applied Research*, *1*(2), 53–57. www.ijsar.in

Sugianto, A. (2017). Validity and Reliability of English Summative Test for Senior High School. *Indonesian EFL Journal: Journal of ELT, Linguistics, and Literature*, *3*(2), 22–38. http://ejournal.kopertais4.or.id/mataraman/index.php/efi

Suwarto. (2011). *Teori Tes Klasik dan Teori Tes Modern*. *1*, 69–78.

Suwarto. (2022). Karakteristik Tes Ilmu Pengetahuan Alam. *Jurnal Pendidikan*, *31*(1), 109. https://doi.org/10.32585/jp.v31i1.2269

Suwarto, S. (2021). *The Characteristics of Indonesia Second-semester Final Test for Eighth-grade Students*. *12*(9), 356–370.

Suzana, A. (2018). Analisis Tingkat Kesukaran dan Daya Beda Butir-Butir Soal Penilaian Akhir Tahun Matematika Kelas X di SMA Negeri 1 Purbalingga. *MathGram Matematika*, *2*(2), 1–8.

Syamsuddin, S., & Istiyono, E. (2018). The effectiveness of mathematics learning through contextual teaching and learning approach in Junior High School. *AIP Conference Proceedings*, *2014*(1), 020085. https://doi.org/10.1063/1.5054489

Syamsuddin, S., & Setiawati, F. A. (2018). The influence of problem solving ability, emotional intelligence and formative tests on learning outcomes of mathematics. *International Conference on Mathematics and Science Education of Universitas Pendidikan Indonesia*, *3*, 803–808. http://science.conference.upi.edu/proceeding/index.php/ICMScE/article/view/174

Syamsuddin, S., & Utami, M. A. P. (2021). Efektivitas Pembelajaran Matematika melalui Pendekatan Contextual Teaching and Learning. *Jurnal Riset Dan Inovasi Pembelajaran*, *1*(1), 32–40. https://doi.org/10.51574/JRIP.V1I1.14)

Uddin, I., Uddin, I., Rehman, I. U., Siyar, M., & Mehboob, U. (2020). Item Analysis of Multiple Choice Questions in Pharmacology. *Journal of Saidu Medical College, Swat*, *10*(2). https://doi.org/10.52206/jsmc.2020.10.2.320

Utami, M. A. P., & Syamsuddin, S. (2022). *An Implementation to Determine The KKM of Music*. *1*(4), 541–549.

Utami, & Syamsuddin. (2020). Perubahan Perilaku Nomophobia melalui Pendekatan Interaksi Sosial: Sngle Case Research (SCR). *Preschool: Jurnal Perkembangan Dan*

Pendidikan Anak Usia Dini, *2*(1), 133–140. http://ejournal.uin-malang.ac.id/index.php/preschool/article/view/10307