COMPETITIVE, Vol. 9, No.1, Juni 2014
ISSN: 0216-2539

Teori Pengukuran Pendidikan Menggunakan *Classical Test Theory* Dan *Item Response Theory*

Muhammad Ruslan Maulani[1], Budi Rahardjo[2]
[1]Jurusan Teknik Informatika Politeknik Pos Indonesia
e-mail: mruslanmaulani@poltekpos.ac.id
Jl. Sariasih No. 54 Bandung, telp: +62-22-200 9570
[2]Program Magister Informatika Institut Teknologi Bandung
e-mail: rahard@gmail.com
Jl. Ganesha No. 10 Bandung, telp: +62-22-250 0935

Abstrak

Makalah ini membahas mengenai teori pengukuran pendidikan. Di dalam pengukuran pendidikan terdapat dua metode yang biasa digunakan yaitu metode pengukuran klasik (Classical Test Theory) dan metode pengukuran modern biasa dikenal dengan metode Item Response Theory.Tujuan dari pembuatan makalah ini yaitu untuk mengetahui teori yang digunakan dalam pengukuran pendidikan serta mengetahui perbedaan antara Classical Test Theory dan Item Response Theory.

Kata kunci: *Classical Test Theory*, *Item Response Theori*, *Pengukuran Pendidikan*

Abstract

This paper discusses the theory of educational measurement. Inside there are two methods of educational measurement that is commonly used is the classical measurement method (Classical Test Theory) and modern measurement methods commonly known as Item Response Theory methods.The objective of this paper is to determine the theory in educational measurement and to know the difference between Classical Test Theory and Item Response Theory.

Keywords: *Classical Test Theory, Item Response Theori, educational measurement*

1. Pendahuluan

Perkembangan teknologi telah membawa perubahan dalam kehidupan manusia. Perubahan tersebut dapat kita rasakan dalam dunia bisnis, pemerintahan, maupun dalam dunia pendidikan. Salah satu dampak dari teknologi dalam dunia pendidikan yaitu terjadinya perubahan dalam proses penyampaian ilmu pengetahuan. Proses penyampaian ilmu pengetahuan kepada pelajar pada awalnya disampaikan melalui tatap muka, tetapi kini telah mengalami perubahan. Proses pembelajaran tersebut dapat juga disampaikan dengan melalui teknologi internet. Proses pembelajaran dengan melalui teknologi internet tersebut dikenal dengan istilah e-learning.

Dalam sistem e-learning, evaluasi terhadap suatu soal merupakan salah satu komponen penting dan merupakan tahap yang harus ditempuh oleh pengajar agar dapat mengetahui efektifitas suatu pembelajaran[1]. Sebenarnya, dengan melakukan penilaian terhadap suatu soal tidak hanya akan menambah efektifitas pembelajaran dan keterampilan siswa, tetapi akan memberikan umpan balik kepada instruktur, sehingga dapat meningkatkan produk dan layanan pembelajaran, serta menentukan strategi dalam proses pembelajaran[2],[3], dan[4].

Untuk mengukur kualitas pertanyaan pada setiap butir soal dapat diketahui melalui parameter tingkat kesulitan (*difficulty)*, fungsi pembeda soal (*discrimination)* dan menjawab dengan cara ditebak (*guessing)*[5]. Dalam teori pengukuran, terdapat dua model pengukuran, yaitu *Classical Test Theory* dan *Item Response Theory*[6].

2. Classical Test Theory (CTT)

Classical Test Theory (CTT) dikembangkan sekitar tahun 1920-an[7]. Teori ini memiliki beberapa komponen seperti teori validitas, reabilitas, objektivitas, teori analisis tes, teori analisis butir dan sebagainya. *Classical Test Theory*adalah suatu model pengukuran berdasarkan informasi yang didapatkan pada level skor test[8]. *Classical Test Theory*adalah teori mengenai skor tes yang mengenalkan tiga konsep yaitu *test score* / *observed score, true score* dan *error score*[9].

CTT didasarkan pada model aditif, yaitu skor amatan merupakan penjumlahan dari skor sebenarnya dengan skor kesalahan pengukuran [10]. Secara matematis pernyataan tersebut dapat dirumuskan sebagai berikut:

$$X = T + E \quad \text{..... (1)}$$

Keterangan:
X = Skor amatan
T = Skor murni
E = skor kesalahan pengukuran (*error score*)

Dalam melakukan pengukuran dengan menggunakan CTT ada beberapa parameter yang dapat digunakan yaitu: tingkat kesukaran, daya pembeda butir soal dan kualitas pengecoh.

2.2. Tingkat Kesukaran

Di dalam kerangka uji tes, terdapat salah satu ciri butir yang dikenal dengan tingkat kesukaran butir. Makin sedikit peserta ujian yang dapat menjawab suatu butir dengan benar, maka makin sukar butir itu. Dan dengan sendirinya, makin banyak peserta uji tes yang dapat menjawab dengan benar suatu butir, maka makin tidak sukar atau makin mudah butir itu.

Secara matematis tingkat kesukaran yang dihitung dengan proporsi menjawab benar dirumuskan dengan:

$$p = \frac{\sum B}{N} \quad \text{..... (2)}$$

Adapun klasifikasi tingkat kesukaran soal dapat dicontohkan sebagai berikut[11]:

- 0,00 – 0,30 : soal tergolong sukar
- 0,31 – 0,70 : soal tergolong sedang
- 0,71 – 1,00 : soal tergolong mudah

Namun angka tersebut di atas perlu disesuaikan dengan tujuan pengembangan soal. Soal untuk keperluan seleksi, remedy atau ulangan umum seharusnya mempunyai tingkat kesukaran yang berbeda agar tercapai tujuan yang maksimal[12].

2.3. Daya Pembeda

Daya pembeda merupakan parameter butir soal yang memberikan informasi tentang seberapa besar butir soal tersebut dapat membedakan peserta tes yang skornya tinggi dengan peserta tes yang skornya rendah[12]. Semakin tinggi daya pembeda suatu butir soal, maka makin besar perbedaan sekor yang dihasilkan oleh kelompok tinggi dan kelompok rendah. Dengan kata lain, makin tinggi daya pembeda butir, makin banyak peserta dari kelompok tinggi yang dapat menjawab butir itu dengan benar serta makin sedikit peserta dari kelompok rendah yang dapat menjawabnya dengan benar. Karena itu, daya pembeda butir melibatkan pembagian peserta ke kelompok tinggi dan kelompok rendah.

Daya pembeda butir dinyatakan dalam bentuk indeks. Indeks tersebut dihitung melalu suatu rumus tertentu. Anastasi mengemukakan bahwa selama ini telah ada lebih dari 50 macam indeks daya pembeda melalui lebih dari 50 macam rumus.

Indeks jenis pertama. Pada indeks jenis pertama ini, indeks dilihat berdasarkan pada kelompok sekor jawaban peserta pada kelompok tinggi dan kelompok rendah. Dalam hal ini, indeks daya pembeda butir ditentukan oleh selisih proporsi jawaban benar pada kelompok tinggi dan kelompok rendah. Makin besar indeks daya pembeda butir, makin besar selisih proporsi jawaban benar di antara kelompok tinggi dan kelompok rendah. Untuk menentukan daya pembeda tersebut dapat menggunakan rumus sebagai berikut:

$$D = \frac{1}{M_T} f_{iT} - \frac{1}{M_R} f_{iR} \text{.....(3)}$$

Keterangan:
D = indeks daya pembeda
M_T= Jumlah peserta pada kelompok tinggi
M_R= Jumlah peserta pada kelompok rendah
f_{iT}= frekuensi jawaban benar pada kelompok tinggi
f_{iR}= frekuensi jawaban benar pada kelompok rendah

Interval daya pembeda butir dapat dilihat pada tabel di bawah.

Tabel 2.1. Interval Daya Pembeda Butir

Interval	Klasifikasi	Interpretasi
a <= 0,20	Jelek	Daya pembeda jelek
0,20 <= a <= 0,40	Memuaskan	Memiliki daya pembeda yang cukup
0,41 <= a <= 0,70	Baik	Memiliki daya pembeda yang baik
0,71 <= a <= 1,00	Sangat baik	Memiliki daya pembeda yang sangat baik

2.4. Kualitas Pengecoh

Soal pilihan ganda sebaiknya memiliki pilihan yang berfungsi sebagai pengecoh, dimana pengecoh tersebut sebenarnya bukan merupakan jawaban yang benar. Setiap pengecoh perlu dibuat sedemikian rupa sehingga menarik perhatian peserta tes yang belum memiliki konsep yang baik terhadap materi yang diujikan. Pengecoh yang baik minimum berindeks 0,1 yang berupa koefisien korelasi point biserial, bernilai positif untuk kunci jawaban dan bernilai negatif untuk pengecoh[10].

CTT telah digunakan selama bertahun – tahun untuk menentukan tingkat kesukaran dan karakteristik lainnya dalam instrument pengukuran[13]. Namun, ada beberapa kekurangan dalam CTT. Kekurangan yang paling menonjol dalam teori ini adalah index butir soal seperti tingkat kesukaran dan daya pembeda yang didapatkan dengan menggunakan CTT bergantung pada kelompok peserta tes dan penilaian kemampuan peserta tes bergantung pada pemilihan butir soal. Untuk mengatasi masalah ini, para ahli psikometri mengembangkan teori pengukuran baru yang disebut *item response theory* (IRT).

2. Item Response Theory(IRT)

Untuk mengatasi kelemahan-kelemahan yang ada pada teori klasik, para ahli pengukuran berusaha untuk mencari alternative. Model yang diinginkan harus mempunyai sifat-sifat : (1) karakteristik butir tidak tergantung kepada kelompok peserta tes yang dikenai butir soal tersebut, (2) skor yang menyatakan kemampuan peserta tes tidak tergantung pada tes, (3) model dinyatakan dalam tingkatan (level) butir soal, tidak dalam tingkatan tes dan (5) model menyediakan ukuran yang tepat untuk setiap skor kemampuan [9].

Item Response Theory (IRT) merupakan salah satu cara untuk menilai kelayakan butir dengan membandingkan rerata penampilan butir soal terhadap kemampuan kelompok yang diramalkan oleh model [9]. IRT merupakan teori statistik umum mengenai item soal ujian dan hasil tes serta bagaimana kinerja soal tersebut sesuai dengan kemampuan atau sifat dari soal yang diukur melalui item butir soal pada tes [14]. Tujuan utama dari IRT adalah memberikan kesamaan antara statistik soal dan estimasi kemampuan [9].

IRT memiliki dua prinsip dasar, yaitu [13]: 1) Kinerja siswa dalam mengerjakan test yang dijelaskan oleh beberapa faktor yaitu kemampuan laten (*latent traits*), kinerja tersebut dapat diukur dengan mengetahui nilai dari soal dan 2) hubungan antara kinerja siswa mengenai jawaban peserta tes dengan kemampuan siswa dapat digambarkan oleh fungsi grafik naik disebut juga dengan *Item Characteristic Curve* (ICC). Semakin tinggi tingkat kemampuan, semakin besar peluang jawaban benar dari suatu butir soal.

Pada item response theory, dikenal tiga jenis *mathematical model* yaitu *one-parameter logistic* (1-PL) *model*, *two-parameter logistic* (2-PL) *modeldan three-parameter*

logistic (3-PL) *model*. Perbedaan dari ketiga *mathematical model* tersebut terletak pada jumlah parameter pertanyaan yang digunakan oleh masing – masing *mathematical model*. Perbedaan parameter pertanyaan menimbulkan perbedaan pada hasil estimasi *ability* peserta ujian.

2.1. *One-Parameter Logistic* (1-PL) *Model*

One-Parameter Logistic Model pertama kali diterbitkan oleh Danish matematika Georg Rasch pada tahun 1960 [15]. Oleh karena itu *one-parameter Logistic Model* sering disebut juga sebagai *Rasch Model*. Rasch model memprediksi probabilitas kebenaran (P(θ)) dari masing – masing pertanyaan berdasarkan tingkat estimasi ability dan satu parameter pertanyaan yaitu *difficulty*.

Persamaan untuk mengetahui probabilitas dimana seorang peserta ujian dengan tingkat *ability*θ mampu menjawab pertanyaan satu parameter dengan benar, yaitu sebagai berikut [15]:

$$P(\theta) = \frac{1}{1+e^{-1(\theta-b)}} \quad \ldots\ldots\ldots\ldots\ldots\ldots (4)$$

Keterangan:
$P(\theta)$= probabilitas peserta ujian dengan *ability* = θ menjawab soal dengan benar
θ= tingkat *ability*
b = nilai parameter *difficulty*
e = nilai eksponensial, yaitu 2.718

2.2. *Two-Parameter Logistic* (2-PL) *Model*

Dalam *item response theory*, model matematika standar untuk *ItemCharacteristic Curve* yaitu bentuk kumulatif dari fungsi logistik. Fungsi logistik pertama kali digunakan pada tahun 1844 dan telah banyak digunakan dalam ilmu biologi untuk model pertumbuhan tanaman dan hewan mulai dari lahir sampai jatuh tempo. Pemodelan tersebut pertama kali digunakan sebagai model untuk *item characteristic curve* pada akhir tahun 1950an dan karena kesederhanaanya, maka model tersebut banyak disukai [15].

Persamaan untuk mengetahui probabilitas dimana seorang peserta ujian dengan tingkat ability = θ mampu menjawab pertanyaan dengan benar untuk *two-parameter*, yaitu sebagai berikut:

$$P(\theta) = \frac{1}{1+e^{-L}} = \frac{1}{1+e^{-a(\theta-b)}} \quad \ldots\ldots\ldots\ldots\ldots\ldots (5)$$

Keterangan:
$P(\theta)$= probabilitas peserta ujian dengan *ability* = θ menjawab soal dengan benar
θ= tingkat *ability*
a = nilai parameter *discrimination*
b = nilai parameter *difficulty*
e = nilai eksponensial, yaitu 2.718
L = a(e-b) merupakan suatu deviasi logistic (logit)

2.3. *Three-Parameter Logistic*(3-PL) *Model*

Salah satu fakta ketika peserta ujian mengikuti proses ujian yaitu adanya peserta ujian yang menjawab soal ujian dengan benar melalui tebakan (*guessing*). Dengan demikian, probabilitas respon yang benar akan semakin kecil karena adanya faktor tebakan (*guessing*) [15]. T*wo-parameter model*dimodifikasi dengan memasukkan faktor *guessing* dengan probabilitas respon terhadap jawaban yang benar[16].

Persamaan untuk mengetahui probabilitas dimana seorang peserta ujian dengan tingkat ability = θ mampu menjawab pertanyaan dengan benar untuk *three-parameter*, yaitu sebagai berikut:

$$P(\theta) = c + (1-c)\frac{1}{1+e^{-a(\theta-b)}} \quad \ldots\ldots\ldots\ldots\ldots\ldots (6)$$

Keterangan:
$P(\theta)$= probabilitas peserta ujian dengan *ability* = θ menjawab soal dengan benar

θ= tingkat *ability*
a = nilai parameter *discrimination*
b = nilai parameter *difficulty*
c = nilai parameter *guessing*
e = nilai eksponensial, yaitu 2.718

3. Penutup

*Classical Test Theory*adalah suatu model pengukuran berdasarkan informasi yang didapatkan pada level skor test. Dalam melakukan pengukuran dengan menggunakan CTT ada beberapa parameter yang dapat digunakan yaitu: tingkat kesukaran, daya pembeda butir soal dan kualitas pengecoh.

Pada item response theory, dikenal tiga jenis *mathematical model* yaitu *one-parameter logistic* (1-PL) *model*, *two-parameter logistic* (2-PL) *modeldan three-parameter logistic* (3-PL) *model*. Perbedaan dari ketiga *mathematical model* tersebut terletak pada jumlah parameter pertanyaan yang digunakan oleh masing – masing *mathematical model*. Perbedaan parameter pertanyaan menimbulkan perbedaan pada hasil estimasi *ability* peserta ujian.

4. DAFTAR PUSTAKA

[1] M. DONG, "A Grid-based E-Assessment IRT-Item Bank Platform," *First International Workshop on Education Technology and Computer Science,* pp. 1012-1015, 2009.

[2] L. Giuseppina, "Item Response Theory for Optimal Questionnaire Design," *Journal of e-Learning and Knowledge Society,* vol. 9, no. 3, pp. 79-93, 2013.

[3] C. Lan, S. Graf, K. Lai and Kinshuk, "Enrichment of Peer Assessment with Agent Negotiation," *IEEE Trans. on Learning Technologies,* vol. 4, no. 1, pp. 35-46, 2011.

[4] C. Romero and S. Ventura, "Educational Data Mining: A Review of the State-of-the-Art," *IEEE Transactions on Systems, Man and Cybernetics-Part C: Applications and Reviews,* vol. 40, no. 6, pp. 601-618, 2010.

[5] O. Vozar, "Adaptive Test Question Selection for Web-based Educational System," *Third International Workshop on Semantic Media Adaptation and Personalization,* pp. 164-169, 2008.

[6] T. G. Courville, "An empirical comparison of item response theory and classical test theory item/person statistics," Texas A&M University, Texas, 2004.

[7] V. Natarajan, "An approach to implementing adaptive testing using item response theory both offline and online," in *Proceedings of the 2009 GMAC Conference on Computerized Adaptive Testing*, India, 2009.

[8] L. O'Connor, C. Radcliff and J. Gedeon, "Applying Systems Design and Item Response Theory to the Problem of Measuring Information Literacy Skill," College & Research Libraries, Kent University, 2002.

[9] R. K. Hambleton and H. Swaminathan, Item Response Theory, Boston, MA: Kluwr-Nijhoff, 1985.

[10] M. Allen and W. Yen, Introduction to measurement theory, Belmont: Wadsworth, 1979.

[11] L. R. Aiken, Psychological Testing and Assesment, Boston: Allyn and Bacon, 1994.

[12] S. Hadi, Pengembangan Computerized Adaptive Test Berbasis Web, Yogyakarta: Aswaja Pressindo, 2013.

[13] R. K. Hambleton, H. Swaminathan and J. H. Rogers, Fundamentals of Item Response Theory, London: Sage Publication, 1991.

[14] Y. L. P. Vega, "Application of Item Response Theory (IRT) for the generation of adaptive assessments in an introductory course on object-oriented programming," IEEE, 2012.

[15] F. Baker, "The Basics of item response theory," ERIC Clearing house on Assessment and Evaluation, 2001.

[16] A. Birnbaum, Some latent trait models and their use in inferring an examinee's ability, Part 5 in F.M. Lord and M.R. Novick. Statistical Theories of Mental Test Scores, MA: Addison-Wesley, 1968.