



# Jurnal

## Penelitian dan Evaluasi Pendidikan

- Adi Suryanto, Abdul Gafur, FX. Sudarsono* Model Evaluasi Program Tutorial Tatap Muka Universitas Terbuka
- Ali Hasmy, Suryanto, Kumaidi* Robustness Model-Model Respons Butir terhadap Pelanggaran Asumsi Independensi Lokal Butir
- Amrina Rizta, Zulkardi, Yusuf Hartono* Pengembangan Soal Penalaran Model TIMSS Matematika SMP
- Eni Winaryati, Suyata, Sumarno* Model Evaluasi dalam Supervisi Pembelajaran IPA Berbasis Lima Domain Sains
- Farida Agus Setiawati, Djemari Mardapi, Saifuddin Azwar* Penskalaan Teori Klasik Instrumen *Multiple Intelligences* Tipe Thurstone dan Likert
- Heri Retnawati* Pendeteksian Keberfungsian Butir Pembeda dengan Indeks Volume Sederhana Berdasarkan Teori Respons Butir Multidimensi
- Jien Tirta Rabarja, Trie Hartiti Retnowati* Evaluasi Pelaksanaan Pembelajaran Seni Budaya di SMA Kabupaten Lombok Timur, NTB
- Nuril Huda, Aliyah A. Rasyid, Pujiati Suyata, Sumarno* Pengembangan Model Penelitian Perspektif Gender
- Primardiana Hermilia Wijayati, Suyata, Sumarno* Model Evaluasi Pembelajaran Berbasis Kaizen di Sekolah Menengah Atas
- Rochmiyati* Model *Peer Assessment* pada Pembelajaran Kolaboratif Elaborasi Peta Konsep IPS Terpadu di Sekolah Menengah Pertama
- Septimar Pribatini, Djemari Mardapi, Sutrisno* Pengembangan Model Penilaian Akhlak Peserta Didik Madrasah Aliyah
- Susi Anto, Djemari Mardapi* Komparasi Metode *Standard Setting* untuk Penentuan KKM Mata Pelajaran Matematika Kelas VIII SMP





# Jurnal

## Penelitian dan Evaluasi Pendidikan

### Penerbit

HIMPUNAN EVALUASI PENDIDIKAN INDONESIA

bekerja sama dengan

PROGRAM PASCASARJANA UNIVERSITAS NEGERI YOGYAKARTA

Berdasarkan MOU Nomor 195 B/J.35.17/LK/04

Penanggungjawab : Djemari Mardapi  
Ketua : Moch Alip  
Sekretaris : Heri Retnawati  
Penyunting : Badrun Kartowagiran  
Sudiyatno  
Aman  
Samsul Hadi  
Jailani  
Penyunting Bahasa : Haryadi  
Suhardi  
Sudiyono

**Koordinator:** Nuchron

**Setting:** Rohmat Purwoko

SEMUA TULISAN YANG ADA DALAM JURNAL PENELITIAN DAN EVALUASI PENDIDIKAN  
BUKAN MERUPAKAN CERMINAN SIKAP DAN ATAU PENDAPAT DEWAN REDAKSI  
TANGGUNGJAWAB TERHADAP ISI DAN ATAU AKIBAT DARI TULISAN  
TETAP TERLETAK PADA PENULIS



## PENSKALAAAN TEORI KLASIK INSTRUMEN *MULTIPLE INTELLIGENCES* TIPE THURSTONE DAN LIKERT

<sup>1)</sup>Farida Agus Setiawati, <sup>2)</sup>Djemari Mardapi, <sup>3)</sup>Saifuddin Azwar

<sup>1,2)</sup>Universitas Negeri Yogyakarta, <sup>3)</sup>Universitas Gajah Mada

<sup>1)</sup>faridaagus@yahoo.co.id, <sup>2)</sup>djemarimardapi@yahoo.co.id, <sup>3)</sup>sfazwar@ugm.ac.id.

### Abstrak

Penelitian ini bertujuan untuk: 1) mengetahui hasil penskalaan instrumen *multiple intelligences* (MI) pada tipe Thurstone dan Likert dengan pendekatan klasik, 2) mengetahui karakteristik instrumen MI pada tipe Thurstone dan Likert pada data asli dan data yang diskalakan, 3) membandingkan karakteristik psikometrik pada kedua tipe data yang sudah diskalakan. Penelitian ini menggunakan pendekatan kuantitatif yang pelaksanaannya terdiri dari empat bagian yang saling terkait, yaitu penelitian pengembangan instrumen, penskalaan pada data hasil ujicoba, analisis karakteristik psikometrik instrumen, dan perbandingan karakteristik psikometrik instrumen. Instrumen dikembangkan menggunakan tipe Thurstone dan Likert pada konstruk yang sama. Perbandingan karakteristik psikometrik kedua instrumen dilakukan secara diskriptif. Hasil penskalaan dengan metode *paired comparison* didapatkan urutan skor stimulus dari yang terendah yaitu: logika matematika, musik, linguistik, kinestetik, naturalis, visual, interpersonal, eksistensial dan intrapersonal. Penskalaan dengan metode *summated rating* dihasilkan skor terstandar dari yang rendah hingga tinggi pada tiap respons. Terdapat perubahan skor, varian, reliabilitas dan kesalahan baku pengukuran (SEM) dari data asli dengan data yang diskalakan. Koefisien reliabilitas dan SEM instrumen tipe Thurstone lebih rendah dibanding tipe Likert.

**Kata kunci:** penskalaan, *multiple intelligences*, tipe Thurstone, tipe Likert

## SCALING CLASSICAL THEORY OF MULTIPLE INTELLIGENCES CLASSICAL INSTRUMENT TYPE THURSTONE AND LIKERT

### Abstract

The study aimed to: 1) result the scaling data of multiple intelligence (MI) instruments of Thurstone and Likert types using the classical approach, 2) reveal the psychometric characteristics of Thurstone and Likert types in the original data and the scaled data, 3) compare the psychometric characteristics of the two types of data. The study used the quantitative research approach. The activity consisted of: developing instruments, processing the data scaling, analyzing the psychometric characteristics of the instruments, and comparing the psychometric characteristics of them. The instrument was developed using Thurstone and Likert types in the same constructs. The comparison of psychometric characteristics of two types of data was analyzed by descriptive statistic. The result of scaling using paired comparison method are the sequential scores from a low to high on mathematical-logical, musical, linguistic, kinesthetic, natural, visual, interpersonal, existential and intrapersonal intelligence. The scaling using summated rating produce scores that vary in each response. There are changes of variants and standard error of measurement (SEM) after transformed data. The reliability and SEM of the Thurstone type are lower than that of Likert type.

**Keywords:** scaling, *multiple intelligence instrument*, Thurstone type, Likert type



## Pendahuluan

Pengukuran aspek nonkognitif ditandai dengan memberi skor atau angka pada hasil pengukuran. Skor menunjukkan kualifikasi atribut yang diukur. Makna skor hasil pengukuran dapat dibedakan dari *level of measurement* atau level pengukuran (Lord & Novick, 1968, pp.20-21, Allen & Yen, 1979, pp.7-9, Cohen & Swerdlik, 2009, pp.73-74). Ada empat level pengukuran, yaitu nominal, ordinal, interval dan rasio. Skor hasil pengukuran instrumen tipe Thurstone dan Likert merupakan data ordinal. Analisis data ordinal, hanya dapat menggunakan modus dan median dan tidak dapat menggunakan mean dan SD. Keterbatasan pengukuran dengan menggunakan data tersebut tidak memenuhi asumsi para-metrik yang terkait dengan data. Glenberg (1988, p.144) menyatakan bahwa asumsi dalam analisis parametrik adalah data yang diperoleh berada pada level interval dan rasio. Dengan demikian data pada level ordinal tidak dapat dianalisis, dengan statistik parametrik dan analisis yang dilakukan adalah nonparametrik.

Penggunaan data ordinal dalam analisis data parametrik menimbulkan perdebatan panjang dari para ahli. Hingga saat ini hal itu belum mendapatkan kesepakatan. James Carifio & Rocco Perla (2008, p.1150) menyatakan hal yang sama terkait dengan adanya perdebatan panjang hingga 50 tahun tentang data yang didapatkan pada model pengukuran Likert. Jamieson (2004, p.1212) memberi informasi bahwa karakteristik data pada skala Likert merupakan karakteristik data ordinal, atau *rank order data*, sehingga harus dianalisis dengan nonparametrik, yang kurang sensitif dan kurang *powerful* dibanding parametrik. Suryabrata, (1998, p.146) menyatakan bahwa data yang diperoleh dalam pengukuran psikologi belum benar-benar data interval, tetapi diperlakukan sebagai data interval. Salah satu upaya untuk membuat data menjadi data interval pada hasil pengukuran psikologi adalah dilakukannya proses penskalaan, sehingga proses penskalaan pada data ordinal dapat menjadi

salah satu alternatif pemecahan perdebatan panjang tersebut.

Proses penskalaan merupakan upaya untuk menempatkan atribut atau karakteristik pada suatu rentang kontinum, yang didalamnya melibatkan perubahan nilai atau transformasi skor baik berupa *transformasi linear* maupun *nonlinear* (Brennan, 2006, p.155). Dalam penelitian ini penskalaan dikaitkan dengan upaya untuk menempatkan atribut psikologi dengan mengubah atau mentransformasi data yang semula berbentuk data ordinal yang tidak memiliki unit pembandingan yang sama menjadi data interval atau rasio yang memiliki satuan pembandingan yang sama.

Berbagai macam cara atau metode dilakukan dalam upaya untuk mentransformasi data menjadi data interval. Metode penskalaan tidak terlepas dari pendekatan yang digunakan. Terkait dengan pengukuran aspek psikologi, Torgerson (1958, p.46) mengelompokkan metode penskalaan dalam tiga metode, yaitu: (1) metode penskalaan yang berpusat pada stimulus; (2) metode penskalaan yang berpusat pada respons; dan (3) metode penskalaan yang berpusat pada subjek.

Metode penskalaan berbasis stimulus merupakan metode penskalaan dengan menempatkan serangkaian stimulus dalam suatu kontinum poin. Metode penskalaan berpusat pada respons merupakan metode penskalaan dimana respons subjek dibuat dalam suatu rentang poin tertentu. Metode ini disusun berdasar distribusi respons subjek. Salah satu contoh metode ini adalah skala Likert. Respons subjek diberikan dalam taraf persetujuan atau ketidaksetujuan dalam berbagai variasinya. Pada metode penskalaan yang berpusat pada subjek, penyusunan tes meletakkan subjek atau individu yang akan dihadapi pada poin yang berbeda secara kontinum. Metode ini banyak digunakan, misalnya dalam pengelompokan siswa menggunakan skor z, persentil, skor IQ.

Pada 10 tahun terakhir ini, pengukuran aspek psikis manusia kembali berfokus pada metode yang awalnya sudah ber-



kembang sejak tahun 1930-an, yaitu metode yang rentan untuk dijawab tidak jujur atau *faking* (Chernyshenko, *et.al.*, 2009, p.106). Salah satu metode yang digunakan adalah perbandingan berpasangan yang dipelopori oleh Thurstone (1927, p.383). Dalam metode ini disajikan 2 stimulus pada tiap butir untuk dipilih yang paling sesuai dengan keadaan subjek. Metode ini selanjutnya menjadi dasar pengukuran *force-choice*, yaitu metode pengukuran yang mengharuskan subjek memilih satu pernyataan dari beberapa pernyataan yang disediakan (McDonald, 1999, p.24). Pemilihan satu dari beberapa pernyataan dalam *paired comparison* menyebabkan tipe ini memiliki karakteristik yang berbeda dengan tipe instrumen yang menyajikan sebuah pernyataan dengan berbagai respons.

Penskalaan pada tipe ini berbasis stimulus. Dalam metode ini serangkaian stimulus diperbandingkan untuk direspons sesuai dengan kondisi subjek. Dengan demikian, tipe instrumen ini memberikan alternatif pada responden memberikan respons yang berbeda-beda pada berbagai stimulus yang disajikan, sehingga akan didapatkan respons yang bervariasi pada berbagai stimulus yang disajikan. Hal ini sejalan dengan pendapat Olivares & Brown (2010, p.935) yang menyebutkan bahwa tipe ini terhindar dari jawaban yang sama atau adanya bias dalam memberi respons seperti respons persetujuan yang ekstrim, atau kelemahannya dalam pemberian respons yang tidak bervariasi atau *halo-effect*.

Instrumen tipe *force-choice* memiliki kelebihan terkait dengan respons subjek yang cenderung terhindar dari *social desirability* dan *faking* (McDonald, 1999, p.24, Chernyshenko, *et. al.*, 2009, p.108). *Social desirability* yaitu pernyataan bersifat umum yang memiliki kecenderungan untuk direspons subjek dengan tidak jujur, sehingga sulit untuk mengetahui pendapat subjek sebenarnya. Sementara itu *faking* adalah kecenderungan subjek memilih respons yang tidak sesuai dengan karakteristik dirinya atau dengan sengaja mengubah responsnya untuk men-

dapatkan dirinya pada kelompok yang diharapkan.

Metode *paired comparison* memiliki kelemahan terkait dengan beban kerjanya yang banyak akibat dari setiap butir yang harus mendapat kesempatan untuk berpasangan dengan butir dari aspek lain (Kwan & Chiu, 2007, p.433). Kelemahan lain model ini menurut McIver & Carmines (1986, p. 21) adalah dipengaruhi oleh karakteristik *judgment*. Hal ini disebabkan karena proses penskalaan model ini terkait dengan orang yang akan memberikan *judgment*, maka karakteristik *judgment* juga akan mempengaruhi hasil penskalaan. Berbagai kelemahan dalam tipe ini dapat diatasi dengan model pengukuran yang berbasis respons subjek. Metode yang sering digunakan dan sangat terkenal dalam pengembangan instrumen pada penelitian sosial adalah metode yang digunakan dalam instrumen tipe Likert.

Instrumen tipe Likert muncul dalam jurnal *Archives of Psychology* yang berjudul *Technique for Measurement of Attitudes*. Penskalaan tipe Likert dikembangkan oleh Rensis Linkert untuk mengukur sikap sosial. Pengukuran tipe ini pertama kali dibuat untuk mengukur sikap atau pendapat seseorang terhadap hubungan internasional yang terkait di dalamnya adalah berbagai opini tentang hubungan Amerika terhadap perdamaian dan konflik dengan negara lain, sikap terhadap kegiatan militer Amerika dan sikap terhadap orang kulit hitam atau Negro (Likert, 1932, pp.15-20). Meskipun instrumen ini digunakan untuk mengukur sikap, namun lebih lanjut ia juga menyatakan bahwa tipe pengukuran ini tidak hanya dapat digunakan untuk mengukur pendapat atau sikap tetapi juga digunakan untuk mengukur persepsi dan berbagai karakteristik psikis atau *trait* manusia serta dalam pengukuran kepribadian (Likert, 1932, p.7). Davies (2008, p.134) juga menyatakan bahwa instrumen tipe Likert ini berkembang luas dan digunakan tidak hanya dalam mengukur sikap dan opini tetapi juga mengukur performansi dan kemampuan manusia.

Model ini disusun berdasar distribusi respons subjek yang berbeda dengan model



Thurstone dalam proses penskalaan. Dalam tipe ini respons subjek diberikan dalam taraf persetujuan, misalkan: sangat setuju atau *strongly approve*, setuju atau *approve*, tidak tentu atau *undecided*, tidak setuju atau *disapprove*, dan sangat tidak setuju atau *strongly disapprove* (Likert, 1932, p.14). Dalam perkembangan selanjutnya, respons subjek ini tidak hanya berbentuk taraf persetujuan, tetapi berkembang dalam berbagai format, misalnya: taraf kesesuaian, taraf keyakinan, frekuensi, ketertarikan dan skala *rating* lainnya yang menggambarkan aktivitas, perasaan atau situasi tertentu (Massof, 2004, p. 382). Jumlah respons dalam tipe ini juga berkembang dan bervariasi, yang semula tipe ini menggunakan 5 respons, berkembang menjadi 7 respons (Van Zanten, *et.al.*, 2006, p. 521). Proses penskalaan tipe Likert dikenakan pada respons subjek dengan berbagai variasinya, atau kriteria yang digunakan dalam penskalaan adalah respons, maka penskalaan model ini dapat dikatakan sebagai proses penskalaan dengan pendekatan respons.

Tipe Likert memiliki beberapa kelebihan dan kelemahan. Kelebihan tipe Likert terkait dengan kesederhanaannya dan kemudahannya dalam penyusunan dan interpretasi instrumen (Laerhoven, Zaag-Loonen, Derkx, 2004, p. 830). Namun, kritik terhadap tipe ini adalah adanya keterkaitan atau korelasi yang tinggi antara satu butir dengan butir-butir lain atau antara butir dengan total. Korelasi tinggi ini akan berpengaruh pada skor kekuatan butir yang digunakan dalam pertimbangan seleksi butir (McIver & Carmines, 1986, p.30). Kelemahan lain dari tipe ini adalah memungkinkan responden untuk melakukan *faking*. Keberadaan *faking* ini memungkinkan diperoleh skor instrumen yang tidak sesuai dengan kondisi sebenarnya.

Penskalaan merupakan bagian yang mendasar dalam proses pembentukan teori pengukuran (Lord & Novick, 1968, p.22). Penskalaan dengan teori klasik yang terkenal dilakukan oleh Thurstone dan Likert. Thurstone melakukan proses penskalaan dengan menggunakan metode *paired comparison* de-

ngan mengabaikan asumsi distribusi variasi subjek atau case V (Guilford, 1954, p.156). Sementara itu, Likert (1923, p.21) menggunakan metode Sigma dalam proses penskalaannya. Dalam metode ini setiap respons pada setiap pernyataan diberi skor dengan berasumsi pada distribusi normal. Penggunaan distribusi normal ini dilakukan agar setiap skor dapat diperbandingkan. Selanjutnya metode ini disebut juga dengan *summated rating* (Shaw & Jack, 1967, p.24, Azwar, 2004, p.104, Dunn-Rankin, *et.al.*, 2004, p.105). Proses penskalaan Thurstone dan Likert dilakukan dengan dengan prinsip yang sama yaitu mengkonversi skor kasar yang didapatkan dengan skor z yang menggunakan distribusi normal. Dengan proses penskalaan ini jarak antarskor menggunakan satuan yang sama, sehingga data yang didapatkan menjadi data interval.

Pengembangan instrumen ilmu-ilmu sosial terutama psikologi banyak menggunakan tipe Thurstone ataupun Likert. Begitu pula dengan instrumen *multiple intelligences* (MI) yang dikembangkan dalam penelitian disertasi ini. Di beberapa negara instrumen *multiple intelligences* sudah dibuat dan dikembangkan dalam berbagai penelitian. Pada tahun 2007, peneliti juga sudah mengembangkan instrumen *multiple intelligences*, namun instrumen yang sudah dikembangkan ini dalam analisisnya belum dilakukan proses penskalaan. Di samping itu, terdapat beberapa butir yang muatan faktor yang rendah sehingga perlu diperbaiki. Dengan demikian, perlu dilakukan penelitian lebih lanjut untuk mengembangkan instrumen yang sudah dibuat dengan melakukan proses penskalaan pada data hasil penelitian.

Istilah karakteristik psikometrik terkait dengan atribut tes psikologi. Karakteristik psikometrik merupakan berbagai karakteristik yang terkait dengan atribut tentang tes (Furr & Bacharach, 2008, p.8). Berbagai atribut tentang tes psikologi antara lain: tipe data atau skor hasil pengukuran, reliabilitas data hasil pengukuran, dan validitas data hasil pengukuran. Dalam pendekatan teori klasik atribut tes banyak dijelaskan dari konsep tentang reliabilitas.



Konsep reliabilitas muncul dari asumsi dalam teori tes klasik yang terkait dengan skor tampak (X), skor murni (T), dan skor kesalahan pengukuran (E). Konsep utama dalam teori ini adalah  $X = T + E$ , atau skor tampak merupakan gabungan dari skor murni dan kesalahan pengukuran. Berdasar asumsi tersebut, konsep tentang varian skor tampak atau *total variance* yang dihasilkan merupakan gabungan dari varian skor murni atau *true variance* dan varian skor kesalahan atau *error variance*. Asumsi tersebut dibuat dalam formula  $\sigma_x^2 = \sigma_t^2 + \sigma_e^2$  (Gulliksen, 1950, p.34, Lord & Novick, 1968, pp.56-57, Allen & Yen, 1979, p.61, Thissen & Wainer, 2001, pp.26-27).

Reliabilitas terkait dengan kesalahan pengukuran pada data hasil pengukuran. Mardapi (2008, p.32), menyatakan kesalahan pengukuran merupakan komponen ketidakreliabilan yang banyak dikaji dalam ilmu sosial. Berdasar konsep skor dalam teori klasik, reliabilitas data hasil pengukuran dapat dijelaskan dari varian skor. Adanya keterkaitan antara varian skor sebagaimana disebutkan dalam asumsi teori klasik dapat digunakan untuk menjelaskan definisi reliabilitas yang merupakan interaksi varian kesalahan dan varian skor tampak. Konsep tentang reliabilitas dapat diformulasikan sebagai  $\rho_{xx'} = 1 - \sigma_e^2 / \sigma_x^2$ , yang besarnya reliabilitas dipengaruhi oleh varian kesalahan dan varian skor yang tampak. Berdasar formula tersebut, dapat dijelaskan bahwa semakin besar varian kesalahan akan semakin kecil reliabilitas yang dihasilkan dan sebaliknya.

Istilah varian kesalahan sering disebut sebagai *standard error of measurement* (SEM) atau kesalahan baku pengukuran. Formula SEM ini lebih sering dituliskan sebagai SEM atau  $\sigma_e = \sigma_t \sqrt{1 - \rho_{xx'}}$ . Besarnya SEM mempengaruhi reliabilitas maka dapat dikatakan bahwa ketepatan hasil pengukuran dapat dilihat dari besarnya SEM. Semakin kecil SEM pengukuran akan semakin tepat, reliabel dan dapat dipercaya hasil pengukurannya.

Reliabilitas suatu alat ukur diketahui dari koefisien reliabilitas yang diukur

dengan berbagai metode atau cara. Secara umum ada 3 cara untuk mengukur reliabilitas instrumen, yaitu: (1) tes ulang atau *test-retest*, (2) tes paralel atau *parallel form* dan (3) konsistensi internal atau *internal consistency*. Dalam pendekatan tes ulang alat tes diberikan pada sekelompok subjek dua kali, dengan selang waktu tertentu, misalnya selang dua minggu. Koefisien reliabilitas diukur dari korelasi skor pada tes pertama dan tes kedua. Korelasi antarskor pada tes pertama dan tes kedua dapat dihitung dengan rumus korelasi *product moment*. Pendekatan tes ulang mengandung kelemahan karena kondisi subjek pada tes pertama dan kedua bisa berubah karena unsur belajar, pengalaman dan motivasi yang berbeda. Meskipun demikian, pendekatan ini cocok digunakan untuk pengukuran-pengukuran keterampilan, terutama keterampilan fisik.

Reliabilitas pada pendekatan tes paralel didapatkan dengan cara memberikan pengukuran dengan menggunakan dua perangkat tes yang paralel, misalnya perangkat pertama dan kedua diberikan pada sekelompok subjek. Koefisien reliabilitas tes diukur dengan menghitung korelasi antara skor perangkat pertama dan kedua. Koefisien korelasi juga dapat diukur dengan menggunakan korelasi *product moment*. Metode ini memiliki keterbatasan karena sulitnya menyusun perangkat tes yang paralel.

Pada pendekatan konsistensi internal, pengukuran diberikan pada sekelompok subjek sekali, lalu dengan cara tertentu dihitung koefisien reliabilitas tes tersebut. Ada berbagai formula yang digunakan dalam menghitung reliabilitas. Salah satu cara menghitung koefisien reliabilitas dilakukan dengan menganalisis varian skor dengan menggunakan formula *alpha* dari Cronbach. Perhitungan reliabilitas dengan menggunakan koefisien *alpha* lebih banyak digunakan dibanding perhitungan dengan teknik lain. Teknik ini memiliki kelebihan pada kemudahannya dalam mendapatkan skor dan cara analisisnya.

Berdasar paparan latar belakang penelitian yang dijelaskan sebelumnya, penelitian ini bertujuan sebagai berikut: (1) menge-



tahui hasil penskalaan instrumen *multiple intelligences* pada tipe Thurstone dan Likert dengan pendekatan klasik; (2) mengetahui karakteristik instrumen *multiple intelligences* pada tipe Thurstone dan Likert pada data asli dan data yang diskalakan; (3) membandingkan karakteristik psikometrik pada kedua tipe data yang sudah diskalakan dengan menggunakan pendekatan klasik

### Metode Penelitian

Secara umum penelitian ini menggunakan pendekatan penelitian kuantitatif yang dalam pelaksanaannya terdiri dari beberapa bagian penelitian yang saling terkait, yaitu: penelitian pengembangan instrumen, proses penskalaan pada data hasil uji coba, analisis karakteristik psikometrik instrumen, dan perbandingan karakteristik psikometrik instrumen.

Bagian pertama adalah penelitian pengembangan. Pada penelitian pengembangan dibuat instrumen *multiple intelligences* dengan menggunakan dua tipe, yaitu tipe Thurstone dan tipe Likert. Setelah diuji validitas isi dari ahli selanjutnya diperbaiki. Instrumen yang sudah teruji validitas isinya selanjutnya diujicobakan dilapangan. Bagian berikutnya adalah melakukan proses penskalaan. Proses penskalaan dilakukan pada data hasil uji coba di lapangan dengan menggunakan pendekatan klasik. Setelah diskalakan akan dianalisis karakteristik psikometrik kedua tipe instrumen. Bagian terakhir dalam penelitian ini adalah membandingkan karakteristik psikometrik kedua tipe instrumen secara diskriptif.

Subjek dalam penelitian ini adalah mahasiswa Universitas Negeri Yogyakarta (UNY). Pengambilan data dilakukan dengan *two stage cluster sampling* atau pengambilan sample dengan 2 tingkat. Dalam penelitian ini ada 2 tingkatan kluster yaitu tingkat fakultas dan tingkat program studi. Ada 454 mahasiswa yang terlibat dalam penelitian ini yang berasal dari 6 Fakultas dan 12 program studi di UNY. Dari 454 subjek penelitian, terdapat 11 orang yang datanya tidak dapat dianalisis karena tidak lengkap dalam peng-

isian instrumen, sehingga yang data yang berhasil dianalisis sebanyak 443 orang.

Setelah dilakukan proses penskalaan, selanjutnya dianalisis karakteristik kedua tipe instrumen berdasar koefisien reliabilitas. Pada penelitian ini reliabilitas instrumen tipe Thurstone dianalisis dengan dua cara. Reliabilitas hasil pengukuran setiap dimensi dianalisis dengan rumus *alpha* dan reliabilitas data secara keseluruhan dihitung berdasar koefisien reliabilitas internal dari Kendall dan Smith (1940, p.330). Perhitungan reliabilitas instrumen tipe Likert dilakukan dengan rumus *alpha*. Karena instrumen pada penelitian ini memiliki beberapa dimensi, maka analisis reliabilitas menggunakan reliabilitas komposit dari alpha. Perbandingan reliabilitas kedua tipe instrumen dilakukan secara diskriptif.

### Hasil Penelitian dan Pembahasan

Penelitian ini memaparkan proses penskalaan yang dimulai dari menyusun instrumen, mengubah skor hingga menganalisis karakteristik psikometrik hasil penskalaan dengan pendekatan klasik. Penskalaan dilakukan dengan menggunakan instrumen *multiple intelligence* yang butir-butirnya sudah dibuat dan diseleksi, selanjutnya butir-butir yang sudah terseleksi disusun kembali sebagai instrumen dengan format atau tipe Thurstone dan Likert, sehingga kedua tipe instrumen tersebut memiliki konstuk dan butir-butir yang sama namun tipe instrumennya yang berbeda.

Penyusunan butir atau pernyataan instrumen tipe Thurstone mengacu pada metode *paired comparison* atau perbandingan berpasangan. Setiap butir berpasangan dengan butir lain dan dipilih salah satu butir yang paling sesuai. Dalam penelitian ini dipasangkan satu butir pada satu dimensi dipasangkan dengan butir pada dimensi lain. Ada 9 dimensi yang diungkap pada instrumen *multiple intelligence* ini, untuk itu setiap butir pada satu dimensi selalu berpasangan dengan butir pada dimensi yang lain. Oleh karena ada 9 dimensi maka diperlukan 36 pasang butir, sehingga jumlah keseluruhan



ada 72 butir. Setiap dimensi diwakili oleh 8 butir. Penentuan butir-butir yang akan berpasangan dengan butir-butir dimensi lain dilakukan secara random. Pembuatan instrumen tipe Thurstone dengan membuat pasangan butir dengan cara merandom butir setiap pasangan diteliti juga melalui penelitian uji coba dan didapatkan data yang cenderung sama atau identik. Hasil uji coba ini digunakan sebagai dasar penyusunan instrumen tipe Thurstone dalam penelitian ini yang butir-butirnya disusun secara acak pada setiap dimensinya untuk dipasangkan pada butir-butir di dimensi yang lain

Instrumen tipe Likert disusun dengan membuat sejumlah pernyataan yang memiliki konstruk yang sama dengan instrumen tipe Thurstone. Setiap butir disajikan dengan sebuah pernyataan dan 5 alternatif respons yang berada di sampingnya. Alternatif respons dibuat secara urut dari sangat tidak mampu hingga sangat mampu. Sangat tidak mampu diberi skor 1, tidak mampu diskor 2, ragu-ragu diskor 3, mampu diskor 4 dan sangat mampu diskor 5. Oleh karena jumlah butir pada instrumen tipe Likert menyesuaikan dengan instrumen tipe Thurstone, setiap dimensi pada instrumen tipe ini juga diwakili oleh 8 butir, sehingga jumlah keseluruhannya ada 72 butir.

Kedua bentuk instrumen ini selanjutnya diujikan pada mahasiswa UNY yang menjadi subjek penelitian. Data yang didapatkan selanjutnya dianalisis dengan pendekatan klasik. Penskalaan dengan pendekatan klasik dilakukan dengan metode *paired comparison* pada instrumen tipe Thurstone dan metode *summated rating* pada instrumen tipe Likert.

#### Penskalaan *Paired Comparison*

Proses penskalaan dilakukan pada skor hasil penelitian sebanyak 443 subjek. Perhitungan dilakukan secara manual dengan bantuan program *Microsoft Excel* dengan mengikuti langkah-langkah sebagai berikut:

- (1) Membuat matrik frekuensi (f) yang memuat jumlah frekuensi pilihan subjek

terhadap masing-masing pasangan stimulus

- (2) Mengurutkan stimulus dari yang terkecil hingga terbesar
- (3) Mengubah nilai frekwensi menjadi nilai proporsi (p),  $p = f/N$
- (4) Mengkonversikan harga p dalam bentuk skor z yang merupakan deviasi dari rata-rata pada distribusi normal
- (5) Menghitung rata-rata skor z pada tiap dimensi
- (6) Membuat urutan stimulus dari z terkecil hingga besar

Berdasar perhitungan menggunakan langkah-langkah sebagaimana dijelaskan di atas, didapatkan skor hasil transformasi untuk logika matematika 0, musik 0,376, linguistik 0,429, kinestetik 0,575, naturalis 0,668, visual 0,714, interpersonal 1,031, eksistensial 1,065 dan intrapersonal 1,541. Hasil tersebut menunjukkan bahwa dengan menggunakan kriteria skor z, sembilan bentuk kecerdasan di atas, jika diurutkan dalam rentang skor dari terendah hingga tertinggi didapatkan nilai skala yang berbeda beda. Dibanding kecerdasan yang lain, kecerdasan logika matematika merupakan kecerdasan terendah. Hal ini menunjukkan bahwa masih banyaknya mahasiswa yang menganggap soal-soal yang terkait dengan logika matematika lebih sulit dibanding dengan mata pelajaran yang lain.

#### Penskalaan *Summated Rating*

Proses penskalaan instrumen tipe Likert dilakukan dengan metode *summated rating*. Perhitungan penskalaan dilakukan dengan bantuan program *Microsoft Excel*. Perhitungan penskalaan dilakukan dengan menggunakan langkah-langkah sebagai berikut:

- (1) Menghitung jumlah frekuensi (f) respons subjek di tiap kriteria pada tiap butir.
- (2) Skor frekuensi ini selanjutnya diubah menjadi skor proporsi (p) dan proporsi kumulatif. Skor proporsi dihitung dengan cara membagi frekwensi (f) dengan banyaknya responden (N).



- (3) Menghitung Proporsi kumulatif (pk), didapatkan dari proporsi pada tiap kategori ditambah proporsi kategori sebelumnya.
- (4) Proses berikutnya menghitung pk tengah yaitu titik tengah proporsi kumulatif yang dihitung dari setengah proporsi dalam kategori ditambah dengan pk kategori sebelumnya, atau dapat dirumuskan sebagai berikut;  $pk\text{-tengah} = \frac{1}{2}p + pkb$ .
- (5) Proses selanjutnya menghitung nilai deviasi (z) dengan mengkonversi skor pk-tengah menjadi skor z dengan mengacu pada tabel z kurve normal.

Contoh perhitungan penskalaan *summated rating* pada salah satu butir dapat dilihat pada Tabel 1.

Tabel 1. Perhitungan Penskalaan *Summated Rating* pada Salah Satu Butir

Butir 1	Respons				
	1	2	3	4	5
f	4	36	120	229	54
p	0,009	0,081	0,271	0,517	0,122
pk	0,009	0,090	0,361	0,878	1,000
pk tengah	0,005	0,047	0,204	0,541	0,771
z	-2,611	-1,671	-0,826	0,103	0,741
z+	0,000	0,940	1,785	2,714	3,352

Hasil penskalaan *summated rating* didapatkan skor z tiap respons pada tiap butir. Hasil tersebut menunjukkan bahwa dengan proses penskalaan didapatkan skor respons pada tiap butir berbeda dengan skor respons tanpa penskalaan. Skor hasil penskalaan menunjukkan jarak skor antar respons pada masing-masing butir tidak tetap atau sama dengan 1. Skor pada respons 1 berubah menjadi -2,611, respons 2 berubah menjadi -1,671, respons 3 berubah menjadi -0,826, respons 4 berubah menjadi 2,714 dan respons 5 berubah menjadi 0,741. Apabila skor terendah diubah menjadi 0 maka skor 2 berubah menjadi 0,940,

skor 3 berubah menjadi 1,785, skor berubah menjadi 2,714 dan skor 5 berubah menjadi 3,352.

Dengan melihat kedua proses penskalaan instrumen dengan kedua metode yang telah dijelaskan diatas dapat diketahui bahwa penskalaan instrumen tipe Likert dengan metode *summated rating* pada hakikatnya adalah proses penskalaan dengan pendekatan respons, sedangkan penskalaan tipe Thurstone dengan metode *paired comparison* merupakan penskalaan dengan pendekatan stimulus. Dalam aplikasinya instrumen tipe Thurstone ini juga digunakan untuk mengukur karakteristik psikologi seseorang yang dasar pengukuran dan analisisnya menggunakan respons subjek. Sehingga model instrumen tipe Thurstone ini diskalakan juga dengan pendekatan respons. Dalam penelitian ini upaya untuk membandingkan kedua tipe instrumen dengan menganalisis penskalaan instrumen tipe Thurstone dengan menggunakan pendekatan respons. Pada teori klasik kedua instrumen dianalisis dengan menggunakan metode *summated rating*.

#### Deskripsi Data Kemampuan ( $\theta$ ) Subjek Penelitian Sebelum dan Setelah Ditransformasi

Deskripsi data hasil penelitian disajikan dalam dua bentuk, yaitu data sebelum diskalakan yang masih berupa skor asli dan data yang sudah diskalakan atau data yang sudah ditransformasi dalam skor z. Deskripsi data yang disajikan pada setiap dimensi atau bentuk kecerdasan. Data disajikan dalam rata-rata (mean), simpangan baku (SD), median (Md), skor terendah (Min) dan skor tertinggi (Max). Deskripsi data disajikan terpisah setiap bentuk instrumen. Data hasil penelitian dengan menggunakan instrumen tipe Thurstone disajikan pada Tabel 2. dan deskripsi data hasil penelitian dengan menggunakan instrumen tipe Likert disajikan pada Tabel 3.



Tabel 2. Deskripsi Data Penelitian dengan Menggunakan Instrumen Tipe Thurstone

Dimensi	Thurstone									
	Data Asli					Data setelah ditransformasi				
	Mean	SD	Md	Min	Max	Mean	SD	Md	Min	Max
Linguistik	3,255	1,427	3	0	8	3,413	1,415	3,398	0	7,6
Logika-matematika	2,047	1,812	2	0	8	1,746	1,508	1,634	0	6,538
Visual-spasial	3,966	1,625	4	0	8	4,371	1,616	4,482	0	8,229
Musik	2,995	2,11	3	0	8	2,841	1,913	2,62	0	7,236
Kinestetik	3,623	1,497	4	0	8	3,696	1,46	3,646	0	7,775
Intrapersonal	6,296	1,292	6	2	8	8,996	1,773	8,805	2,602	11,3
Interpersonal	4,916	1,306	5	0	8	6,314	1,571	6,489	0	9,499
Naturalis	3,885	1,625	4	0	8	4,121	1,63	4,197	0	8,049
Eksistensial	5,016	1,409	5	1	8	6,184	1,635	6,361	1,13	9,405

Tabel 3. Deskripsi Data Penelitian dengan Menggunakan Instrumen Tipe Likert

Dimensi	Likert									
	Data Asli					Data setelah ditransformasi				
	Mean	SD	Md	Min	Max	Mean	SD	Md	Min	Max
Linguistik	29,84	3,021	30	15	40	30,31	2,869	30,56	15,03	37,9
Logika-matematika	28,36	4,605	29	12	40	30,16	4,068	30,85	14,75	38,47
Visual-spasial	28,84	4,045	29	16	40	30,12	3,4	30,6	18,41	38,12
Musik	26	5,064	26	12	40	30,08	3,949	30,33	18,2	39,33
Kinestetik	29,74	3,293	30	16	40	30,15	2,875	30,34	17,64	37,77
Intrapersonal	32,332	3,437	32	12	40	30,26	3,18	30,52	11,7	36,53
Interpersonal	31,418	3,168	31	11	40	30,27	3,039	30,42	10,12	37,08
Naturalis	29,190	3,993	29	15	40	30,18	3,53	30,41	16,46	38,09
Eksistensial	32,341	3,440	32	9	40	30,41	3,371	30,75	9,279	36,66

Tabel 2 dan 3 menunjukkan karakteristik data pada berbagai dimensi yang ditunjukkan dari skor rata-rata, simpangan baku, median dan skor terendah dan tertinggi. Dimensi intrapersonal merupakan skor rata-rata tertinggi pada instrumen tipe Thurstone, sedangkan pada instrumen tipe Likert rata-rata tertinggi dicapai dimensi kecerdasan eksistensial, yang memiliki selisih yang sangat kecil dengan dimensi intrapersonal. Skor terendah adalah dimensi logika matematika baik pada instrumen tipe Thurstone maupun Likert. Pada kedua tipe instrumen variasi tertinggi dimiliki oleh dimensi kecerdasan musik

Terdapat perubahan kecenderungan sentral dan variasi data setelah data ditransformasi ke dalam skor z. Skor rata-rata dan median pada kedua tipe instrumen mengalami peningkatan setelah data ditransformasi. Simpangan baku juga mengalami

perubahan pada masing-masing dimensi. Simpangan baku pada dimensi linguistik, logika-matematika, visual-spasial, musik dan kinestetik mengalami penurunan pada instrumen tipe Thurstone, sedangkan pada dimensi intrapersonal, interpersonal, naturalis dan eksistensial mengalami peningkatan setelah ditransformasi ke dalam skor z. Simpangan baku pada semua dimensi instrumen tipe Likert mengalami penurunan.

#### Analisis Reliabilitas

Salah satu karakteristik psikometrik yang sering digunakan dalam pendekatan teori klasik adalah reliabilitas. Dalam penelitian ini, perhitungan reliabilitas instrumen pada tipe Thurstone dihitung dengan rumus koefisien reliabilitas internal dari Kendall dan Smith (1940, p.330), sementara perhitungan reliabilitas tiap dimensi yang di-



lakukan dengan rumus *alpha cronbach*. Hasil perhitungan reliabilitas instrumen secara keseluruhan didapatkan koefisien reliabilitas sebesar 0,64. Hasil tersebut menunjukkan instrumen ini cukup reliabel secara internal.

Perhitungan reliabilitas setiap dimensi pada instrumen tipe Thurstone dihitung dengan mengelompokkan butir-butir yang mengukur dimensi yang sama. Hasil perhitungan reliabilitas tiap dimensi dibedakan pada perhitungan reliabilitas pada data asli dan perhitungan reliabilitas pada data yang sudah ditransformasi. Hasil perhitungan reliabilitas kedua data dapat dilihat pada Tabel 4.

Tabel 4. Reliabilitas Tiap Dimensi Instrumen Tipe Thurstone

Dimensi	Data Asli		Data skor yang ditransformasi	
	Reliabilitas	SEM	Reliabilitas	SEM
Linguistik	0,267	1,222	0,265	1,214
Logika-matematika	0,651	1,071	0,638	0,908
Visual-spasial	0,383	1,277	0,344	1,309
Musik	0,7	1,156	0,687	1,070
Kinestetik	0,236	1,308	0,203	1,303
Intrapersonal	0,25	1,12	0,256	1,529
Interpersonal	0,163	1,195	0,181	1,422
Naturalis	0,376	1,283	0,347	1,317
Eksistensial	0,195	1,264	0,203	1,460

Hasil perhitungan reliabilitas pada tiap dimensi dengan menggunakan rumus alpha didapatkan reliabilitas instrumen pada tiap dimensi tidak mengalami perbedaan yang bermakna pada data asli maupun data yang sudah ditransformasi. Hasil tersebut menunjukkan bahwa tidak ada perbedaan reliabilitas pada data asli dan data yang ditransformasi dengan dengan skor z. Namun, apabila dilihat dari kesalahan baku pengukuran (SEM) dari kedua data, SEM pada data yang ditransformasi mengalami penurunan terutama pada dimensi kecerdasan logika matematika dan musik. Pada dimensi linguistik, visual-spasial, kinestetik dan naturalis, terdapat perubahan kesalahan baku pengukuran, namun perubahannya sangat kecil sehingga

tidak terlihat jelas dalam grafik. Sedangkan pada dimensi intrapersonal dan eksistensial tampak jelas mengalami peningkatan. Hasil tersebut menunjukkan pada dimensi instrumen yang memiliki reliabilitas cukup bagus SEMnya mengalami penurunan pada data yang ditransformasi, namun penurunan itu tidak terjadi terutama pada data yang memiliki reliabilitas sangat rendah.

Perhitungan reliabilitas instrumen tipe Likert juga dihitung pada tiap dimensi dengan menggunakan rumus *alpha*. Hasil perhitungan reliabilitas tiap dimensi dapat dilihat pada Tabel 5.

Tabel 5. Reliabilitas Tiap Dimensi Instrumen Tipe Likert

Dimensi	Data Asli		Data skor yang ditransformasi	
	Reliabilitas	SEM	Reliabilitas	SEM
Linguistik	0,700	1,654	0,690	1,598
Logika-matematika	0,887	1,546	0,888	1,359
Visual-spasial	0,783	1,885	0,775	1,613
Musik	0,856	1,925	0,852	1,517
Kinestetik	0,639	1,980	0,636	1,734
Intrapersonal	0,793	1,563	0,775	1,508
Interpersonal	0,749	1,588	0,735	1,565
Naturalis	0,825	1,670	0,817	1,510
Eksistensial	0,861	1,284	0,842	1,338

Reliabilitas tiap dimensi instrumen tipe Likert sebelum dan setelah proses penskalaan juga tidak banyak perubahan, meskipun reliabilitas pada data asli sedikit lebih tinggi dari data yang sudah diskalakan, namun perubahannya sangat kecil dan kurang bermakna. Namun, apabila dilihat dari kesalahan baku pengukuran (SEM), ternyata SEM pada data yang sudah diskalakan cenderung lebih rendah, kecuali pada dimensi eksistensial yang mengalami sedikit peningkatan.

Setelah didapatkan reliabilitas pada tiap dimensi, selanjutnya dihitung reliabilitas instrumen tipe Likert dengan menggunakan rumus reliabilitas komposit alpha berstrata (Cronbach, Schoneman, Mc Kie, 1965, p.293, Widhiarso, 2009, p.42,43). Hasil perhitungan reliabilitas instrumen secara



keseluruhan dengan menggunakan rumus tersebut didapatkan koefisien reliabilitas sebesar 0,939 dengan SEM sebesar 8,422 pada data asli dan reliabilitas sebesar 0,796 dengan SEM 4,594 pada data yang sudah ditransformasi ke dalam skor z.

Hasil analisis reliabilitas skor hasil pengukuran, didapatkan adanya penurunan koefisien reliabilitas setelah data ditransformasi dalam skor z, meskipun penurunan ini sangat kecil sehingga tidak cukup jelas perbedaannya. Namun, setelah skor hasil pengukuran dianalisis lebih lanjut pada SEM, terjadi penurunan SEM pada data yang ditransformasi. Penurunan ini terlihat cukup tinggi terutama pada instrumen yang memiliki reliabilitas yang tinggi.

Hasil tersebut menunjukkan transformasi data dari level ordinal ke level interval tidak meningkatkan skor reliabilitas, tetapi menurunkan SEM. Penggunaan data pada level ordinal dapat mengurangi kesalahan pengukuran apabila dilakukan proses penskalaan menjadi data interval. Hasil penelitian ini sesuai dengan yang pendapat Salkind (2013, p.31) yang menyebutkan bahwa data pada level pengukuran yang lebih tinggi akan meningkatkan ketepatan dan memberikan informasi yang lebih bagus dibanding level di bawahnya.

Instrumen tipe Thurstone memiliki karakteristik yang berbeda dengan instrumen tipe Likert. Meskipun keduanya memiliki konstruk yang sama, karena dibuat dengan format berbeda dan metode pengukuran yang digunakan juga berbeda memberikan hasil reliabilitas dan SEM yang berbeda pula. Meskipun data hasil pengukuran dari kedua tipe instrumen ini berbeda tetapi hasil pengukuran dari kedua tipe instrumen ini saling berhubungan. Hasil perhitungan korelasi kemampuan pada setiap dimensi dengan menggunakan instrumen tipe Thurstone dan Likert dapat dilihat pada Tabel 6.

Pada Tabel 6 terlihat adanya korelasi pada dimensi yang sama dengan tipe instrumen yang berbeda ( $p < 0,01$ ). Korelasi tertinggi terdapat pada dimensi musik, selanjutnya berturut-turut logika-matematika, naturalisme, visual-spasial, kinestetik, linguistik, interpersonal intrapersonal dan paling rendah adalah dimensi eksistensial. Di samping itu, didapatkan pula adanya korelasi yang tinggi pada dimensi yang berbeda, yaitu dimensi eksistensial dengan interpersonal, intrapersonal dengan interpersonal dan intrapersonal dengan kinestetik.

Tabel 6. Hasil Perhitungan Korelasi antardimensi pada Kedua Tipe instrumen

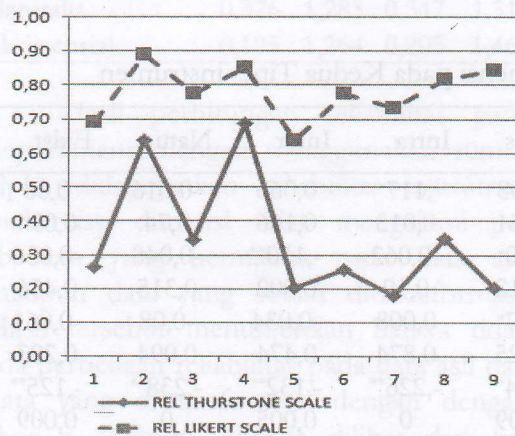
	Ling	Log	Visual	Musik	Kines	Intra	Inter	Natur	Exist
Ling	,311**	-,311**	-,170**	-,146**	-,008	,117*	0,066	-0,016	0,08
	0	0,006	0	0,002	0,861	0,013	0,166	0,74	0,094
Log	0,011	,486**	-,190**	-,216**	,120*	-0,062	-,150**	-0,048	-0,036
	0,821	0	0	0	0,012	0,19	0,002	0,315	0,456
Visual	-0,03	-0,08	,374**	-0,092	,107*	0,008	-0,034	-0,08	-0,061
	0,533	0,094	0	0,052	0,025	0,874	0,474	0,094	0,203
Musik	-,147**	-,156**	0,057	,533**	-,124**	-,226**	-,132**	-,238**	-,125**
	0,002	0,001	0,233	0	0,009	0	0,005	0	0,009
Kines	-,104*	-0,034	-,132**	-0,089	,315**	-,136**	-0,028	-,102*	-,126**
	0,028	0,474	0,005	0,062	0	0,004	0,559	0,032	0,008
Intraps	-,134**	-0,027	0,028	-0,084	,112**	,334**	,123**	0,051	,132**
	0,005	0,573	0,562	0,078	0,018	0	0,009	0,283	0,005
Interps	0,045	-0,093	0,051	0,05	0,067	,126**	,316**	0,028	0,09
	0,347	0,051	0,283	0,297	0,157	0,008	0	0,555	0,058
Natur	-0,07	-0,011	-0,016	,137**	,101*	0,002	0,023	,491**	-0,032
	0,142	0,823	0,741	0,004	0,034	0,966	0,633	0	0,506
Exist	-0,051	-,311**	0,008	-0,014	-0,041	,116*	0,065	0,042	,211**
	0,282	0,005	0,873	0,762	0,385	0,014	0,172	0,379	0



- T : instrumen tipe Thurstone
- L : instrumen tipe Likert
- \*\* : ada korelasi dengan  $p < 0,01$
- \* : ada korelasi dengan  $p$  antara 0,01 hingga 0,5

Adanya korelasi atau hubungan pada dimensi yang sama dengan menggunakan instrumen dengan tipe yang berbeda menunjukkan adanya keterkaitan data hasil pengukuran. Hal ini disebabkan karena data tersebut diambil dari konstruk yang sama, meskipun metode pengukurannya dengan cara berbeda. Demikian pula adanya korelasi yang tinggi pada dimensi yang berbeda, yaitu dimensi eksistensial dengan interpersonal, intrapersonal dengan interpersonal dan intrapersonal dengan kinestetik, menunjukkan bahwa ketiga konstruk dalam instrumen ini memiliki keterkaitan satu dengan yang lain.

Perbandingan reliabilitas kedua tipe instrumen dilakukan secara deskriptif. Berdasar perhitungan reliabilitas yang sudah disajikan pada tabel 4 dan 5 diketahui bahwa pada semua dimensi reliabilitas instrumen tipe Likert lebih tinggi dari instrumen tipe Thurstone, namun lebih jelasnya data reliabilitas kedua tipe instrumen disajikan pada Gambar 1.

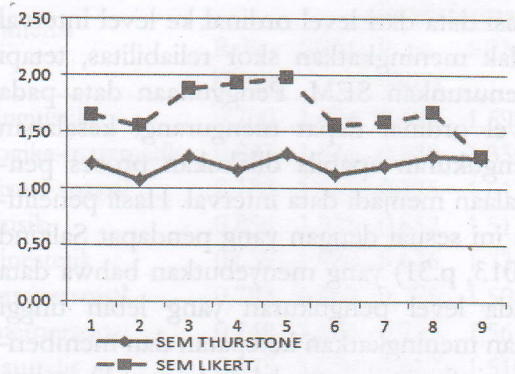


Gambar 1. Reliabilitas Instrumen Tipe Thurstone dan Likert pada Sembilan Kecerdasan

SEM pada skor hasil pengukuran instrumen tipe Thurstone lebih rendah dari SEM pada skor hasil pengukuran dengan instrumen tipe Likert, meskipun reliabilitas

instrumen tipe Likert lebih tinggi dari instrumen tipe Thurstone. Rendahnya varian instrumen tipe Thurstone ini yang menyebabkan reliabilitas dan SEM-nya juga rendah. Dengan demikian, perbedaan yang mendasar pada reliabilitas dan SEM pada kedua tipe instrumen yang dikembangkan dalam penelitian ini lebih disebabkan karena perbedaan varian dari keduanya.

Gambar 1. menunjukkan bahwa terdapat perbedaan reliabilitas yang terlihat jelas pada instrumen tipe Thurstone dan Likert. Reliabilitas instrumen tipe Likert jauh lebih tinggi dari reliabilitas instrumen tipe Thurstone. Meskipun demikian, SEM pada instrumen tipe Thurstone lebih rendah dibanding instrumen tipe Likert. Informasi lebih jelas dapat dilihat pada Gambar 2.



Gambar 2. SEM Instrumen Tipe Thurstone dan Likert pada Sembilan Kecerdasan

Hasil penelitian ini menunjukkan reliabilitas instrumen pada data tipe Thurstone lebih rendah dibanding tipe Likert. Data hasil pengukuran dengan menggunakan instrumen tipe Thurstone lebih rendah konsistennya dibanding tipe Likert. Salah satu faktor yang menyebabkan ketidak-konsistenan hasil pengukuran dengan instrumen tipe Thurstone ini karena setiap butir pada instrumen ini selalu dipasangkan dengan butir lain, sehingga skor sebuah pernyataan akan bergantung dengan seberapa besar pengaruh butir yang menjadi pasangannya. Olivares & Böckenholt (2005, p.265) menyebutkan, untuk mendapatkan respons yang konsisten pada butir-butir yang ber-



pasangan ini diperlukan kondisi yang *transitif*, yaitu sebuah kondisi apabila terdapat 3 variabel A,B,C, dimana  $A > B$ ,  $B > C$ , maka  $A > C$ . Namun, apabila ternyata  $A < C$  maka yang terjadi adalah kondisi yang tidak *transitif* akibat respons yang tidak konsisten secara internal. Kondisi yang *transitif* ini tidak mudah didapatkan apabila variabel yang dibandingkan cukup banyak.

Reliabilitas instrumen berdasar konsistensi internal terkait dengan simpangan baku atau varian data hasil pengukuran. Besar kecilnya reliabilitas terkait dengan variabilitas atau varian skor hasil pengukuran. Begitu pula dalam memaknai dan membandingkan reliabilitas hasil pengukuran, perlu dipertimbangkan pula varian skor pada data yang diperbandingkan. Dengan menggunakan formula SEM atau  $\sigma_e = \sigma_t \sqrt{1 - \rho_{xx}}$  dapat diprediksi besarnya suatu reliabilitas apabila diasumsikan SEM dan varian kedua kelompok sama besarnya.

Dengan mengasumsikan kedua instrumen memiliki SEM yang sama dapat dihitung reliabilitas instrumen yang diprediksi berdasar data pada kelompok sampel yang berbeda atau heterogen. Azwar (1999;72) memberikan persamaan untuk memperoleh prediksi reliabilitas pada data yang diasumsikan memiliki varian yang berbeda sebagaimana disajikan pada formula

$$\rho_{yy} = 1 - \sigma_x^2 (1 - \rho_{xx}) / \sigma_y^2$$

keterangan :

$\rho_{yy}$  = reliabilitas data yang diprediksi

$\sigma_x^2$  = varian pada data yang ditetapkan

$\rho_{xx}$  = reliabilitas data yang sudah ditetapkan

$\sigma_y^2$  = varian data yang akan diprediksi

Dengan menggunakan formula tersebut dapat dihitung prediksi reliabilitas Thurstone berdasar varian sebagaimana data pada instrumen tipe Likert. Hasil perhitungan prediksi reliabilitas kedua data diatas disajikan pada Tabel 7.

Tabel 7. Prediksi Reliabilitas Instrumen Tipe Thurstone apabila Variannya Diasumsikan Sama dengan Tipe Likert

Dimensi	Reliabilitas Semula		Varian		Prediksi reliabilitas Thurstone
	Thurstone	Likert	Thurstone	Likert	
Linguistik	0,265	0,69	2,002	8,231	0,821
Logika-matematika	0,638	0,888	2,274	16,549	0,950
Visual-spasial	0,344	0,775	2,611	11,560	0,852
Musik	0,687	0,852	3,660	15,595	0,926
Kinestetik	0,203	0,636	2,132	8,266	0,794
Intrapersonal	0,256	0,775	3,144	10,112	0,769
Interpersonal	0,181	0,735	2,468	9,236	0,781
Naturalis	0,347	0,817	2,657	12,461	0,861
Eksistensial	0,203	0,842	2,673	11,364	0,812

Tabel 7 menunjukkan bahwa prediksi reliabilitas data pada instrumen tipe Thurstone menjadi tinggi apabila variannya dibuat sama dengan varian data tipe Likert. Kondisi ini disebabkan karena varian data tipe Thurstone yang semula kecil atau rendah menjadi meningkat. Dengan kondisi demikian, rendahnya reliabilitas data semula

pada instrumen tipe Thurstone disebabkan karena variannya yang rendah dan tingginya reliabilitas instrumen tipe Likert disebabkan karena variannya yang tinggi. Lebih lanjut, SEM pada instrumen tipe Thurstone dapat pula diprediksi berdasar varian data pada instrumen tipe Likert, hasil analisis prediksi SEM tersebut dapat dilihat pada Tabel 8.



Tabel 8. Prediksi SEM Tipe Thurstone apabila Varian Datanya Sama dengan Tipe Likert.

Dimensi	Reliabilitas Semula		Varian		SEM Semula		Prediksi reliabilitas Thurstone
	Thurstone	Likert	Thurstone	Likert	Thurstone	Likert	
Linguistik	0,265	0,69	2,002	8,231	1,213	1,597	0,821
Logika-matematika	0,638	0,888	2,274	16,549	0,907	1,361	0,950
Visual-spasial	0,344	0,775	2,611	11,560	1,309	1,613	0,852
Musik	0,687	0,852	3,660	15,595	1,070	1,519	0,926
Kinestetik	0,203	0,636	2,132	8,266	1,304	1,735	0,794
Intrapersonal	0,256	0,775	3,144	10,112	1,529	1,508	0,769
Interpersonal	0,181	0,735	2,468	9,236	1,422	1,564	0,781
Naturalis	0,347	0,817	2,657	12,461	1,317	1,510	0,861
Eksistensial	0,203	0,842	2,673	11,364	1,460	1,340	0,812

Tabel 8 menunjukkan didapatkannya prediksi SEM pada instrumen tipe Thurstone yang semakin tinggi dibanding SEM semula. Meningkatnya SEM pada instrumen tipe Tipe Thurstone ini disebabkan karena variannya disamakan dengan varian tipe Likert. Dengan demikian, rendahnya SEM data semula pada instrumen tipe Thurstone ini dipengaruhi oleh rendahnya varian data tipe Thurstone, dan tingginya SEM pada data tipe Likert juga dipengaruhi oleh lebih tingginya SEM pada data tipe Likert.

Berdasarkan keterangan di atas, besar kecilnya varian turut mempengaruhi besar kecilnya reliabilitas dan SEM. Perbedaan besarnya varian pada kedua tipe instrumen pada penelitian ini mengakibatkan hasil perhitungan SEM pada penelitian ini tampak tidak konsisten. SEM pada skor hasil pengukuran instrumen tipe Thurstone lebih rendah dari SEM pada skor hasil pengukuran dengan instrumen tipe Likert, meskipun reliabilitas instrumen tipe Likert lebih tinggi dari instrumen tipe Thurstone. Rendahnya varian instrumen tipe Thurstone ini yang menyebabkan reliabilitas dan SEMnya juga rendah. Dengan demikian, perbedaan yang mendasar pada reliabilitas dan SEM pada kedua tipe instrumen yang dikembangkan dalam penelitian ini lebih disebabkan karena perbedaan varian dari keduanya.

Ebel & Frisble (1986, p.83) menyatakan bahwa reliabilitas akan lebih tinggi pada kelompok yang memiliki rentang kemampuan yang lebih luas dibanding kelompok yang memiliki rentang kemampuan yang sempit atau homogen. Azwar (1999, p.72) juga menjelaskan bahwa kondisi heterogenitas menyebabkan overestimasi terhadap reliabilitas dan kondisi yang homogen akan menjadi underestimasi. Format instrumen tipe Thurstone memiliki variasi jawaban lebih kecil dibanding tipe Likert. Sedikitnya variasi jawaban instrumen tipe Thurstone ini mengakibatkan simpangan baku dan varian skor hasil pengukuran yang rendah sehingga mempengaruhi pula terhadap perhitungan reliabilitas instrumen dan kesalahan baku pengukuran. Kondisi yang sama juga diteliti oleh Barclay & Wraver (1962, p.119) dengan membandingkan reliabilitas skala sikap tipe Thurstone yang memiliki 2 variasi kategori skor dan Likert 5 variasi kategori. Didapatkan reliabilitas tipe Thurstone sebesar 0,66 dan Likert sebesar 0,97, atau reliabilitas tipe Likert lebih tinggi dari tipe Thurstone.

### Simpulan

Berdasar hasil penelitian yang disajikan pada bab 4, ada beberapa hal yang dapat disimpulkan sebagai berikut.



Penskalaan *multiple intelligence* pada pendekatan klasik didapatkan skor yang telah ditransformasi ke dalam skor z. Dengan menggunakan metode *paired comparison* pada tipe Thurstone didapatkan urutan skor dari yang rendah adalah logika matematika, musik, linguistik, kinestetik, naturalis, visual, interpersonal, eksistensial dan yang paling tinggi adalah intrapersonal. Dengan menggunakan metode *summated rating* pada instrumen tipe Thurstone dan Likert dihasilkan skor z dari yang rendah hingga tinggi pada tiap respons.

Terdapat perubahan varian, reliabilitas dan kesalahan baku pengukuran (SEM) dari data asli dengan data yang ditransformasi menggunakan skor z pada penskalaan teori klasik. Reliabilitas data sebelum dan setelah transformasi tidak terjadi peningkatan, namun SEM pada data yang ditransformasi cenderung mengalami penurunan.

Karakteristik psikometrik dalam bentuk koefisien reliabilitas instrumen tipe Thurstone lebih rendah dibanding tipe Likert, baik pada keseluruhan data maupun data tiap dimensi, begitu pula dalam SEM-nya. SEM pada instrumen tipe Thurstone lebih rendah dibanding tipe Likert. Rendahnya reliabilitas dan SEM pada instrumen tipe Thurstone disebabkan karena variasi tipe Thurstone yang lebih rendah.

#### Saran

1. Proses penskalaan pada teori klasik penelitian ini terbatas menggunakan metode *paired comparison* dan *summated rating*. Penggunaan metode-metode lain seperti metode interval tampak setara, *semantik differential*, dan metode lainnya.
2. Reliabilitas yang rendah pada beberapa dimensi instrumen tipe Thurstone perlu dikaji lagi faktor penyebabnya, sehingga dapat diantisipasi pada pengembangan instrumen.
3. Dalam membahas dan membandingkan konsep reliabilitas dan kesalahan baku pengukuran perlu mempertimbangkan besarnya varian instrumen tersebut.

#### Daftar Pustaka

- Allen, M. J., & Yen, W. M. (1979). *Introduction to measurement theory*. Monterey : Brooks/Cole Publishing Company.
- Azwar, S. (1999). *Dasar-dasar psikometri*. Yogyakarta: Pustaka Pelajar Offset.
- Azwar, S. (2012). *Penyusunan skala psikologis* (Edisi 2). Yogyakarta: Pustaka Pelajar Offset.
- Barclay, J.E. & Weaver, H.B. (1962). Comparative reliabilities and ease of construction of Thurstone and Likert attitude scales. *The Journal of Social Psychology*, 58, 109-120.
- Brennan, R. L. (2006). *Educational measurement* (4<sup>th</sup> Edition). Westport: An Imprint of Greenwood Publishing Group. Inc.
- Carifio, J. & Perla, R. (2008). Resolving the 50-year debate around using and misusing Likert scales. *Medical Education*. 42, 1150-1152.
- Chernyshenko, O. S., Stark, S., Prewett, M., Gray, A.A., Stilson, F. R., & Tuttle, M. D. (2009). Normative scoring of multidimensional pairwise preference personality scales using IRT: empirical comparisons with other formats. *Human Performance*, 22, 105-127.
- Cohen, R. J., & Swerdlik, M. E. (2005). *Psychological testing and assessment, An introduction to tests and measurement* (6<sup>th</sup> Edition). New York: The McGraw-Hill Companies, Inc.
- Cronbach, L. J., Schönemann, P., & McKie, D., (1965). Alpha coefficients for stratified-parallel tests. *Educational and Psychological Measurement*, 25, 291-312.
- Davies, R.S. (2008). Designing a response scale to improve average group response reliability. *Evaluation and Research in Education*, 21, 134.
- Dunn-Rankin, P., Knezek, G. A., Wallace, S. & Zhang, S. (2004). *Scaling methods*, (2<sup>nd</sup> Edition). Mahwah : Lawrence Erlbaum Associates, Inc.



- Ebel, R. L. & Frisbie, D. A. (1986). *Essentials of educational measurement*. New Jersey: Prentice Hall. Inc.
- Furr, R. M., & Bacharach, V. R. (2008). *Psychometrics an introduction*. Los Angeles: Sage Publications
- Glenberg, A. M. (1988). *Learning from data, an introduction to statistical reasoning*. San Diego: Harcourt Brace Jovanovich, Publishers.
- Guilford, J. P. (1954). *Psychometric Methods* (2<sup>nd</sup> Edition). Tokyo: Kōgakusha company, ltd.
- Gulliksen, H., (1950). *Theory of mental tests*. New York: John Wiley and Sons. Inc.
- Jamieson, S. (2004). Likert scales: How to use them. *Medical Education*, 38, 1212
- Kendall, M. G. & Smith, B. B. (1940). On the method of paired comparisons. *Biometrika* 31, 324-345.
- Mardapi, Djemari. (2008). *Teknik penyusunan instrumen tes dan nontes*. Yogyakarta: Mitra Cendekian Offset.
- Kwan, Y.K. & Chiu, L.L. (2007). Modification and replication of Thurstone scaling method and its demonstration with crime seriousness assessment. *The Hong Politecnic University Research Commity*.
- Laerhoven H, van der Zaag-Loonen H.J., & Derkx B.H.F. (2004). A comparison of Likert scale and visual analogue scales as response options in children's questionnaires. *Acta Paediatric*, 93, 830-835.
- Likert, R. (1932). Technique for the measurement of attitudes. New York City: *Archives of Psychology*.
- Lord, F.M. & Novick, M.R. (1974). *Statistical theories of mental test scores*. New York: Addison-Wesley Publishing Company, Inc.
- Massof, R.W. (2004). Likert and Guttman scaling of visual function rating scale questionnaires. *Ophthalmic Epidemiology*, 11, 381-399
- McDonald, R.P. (1999). *Test Theory : A unified threament*. London: Lawrence Erlbaum Associates
- McIver, J.P. & Carmines, E.G. (1986). *Unidimensional scaling*. London: Sage Publications. Inc
- Olivares, A.M. & Boöckenholt, U. (2005). Structural equation modeling of paired-comparison and ranking data. *Psychological Methods*, 10, 285-304
- Salkind, N. J. (2013). *Test and measurement for people who hate tests and measurement*. Los Angeles: SAGE Publication, Inc.
- Shaw, M.E. & Wright, J.M. (1967). *Scales for the measurement of attitudes*. New York: McGraw-Hill Book Company
- Suryabrata, S. (2002). *Pengembangan alat ukur psikologis*. Yogyakarta: Penerbit Andi Offset
- Thissen, & Weiner, D. H. (2001). *Test scoring*. Marwah: Lawrence Erbaum Associated.
- Thurstone, L.L. (1927). A Law of comparative judgment. *Psychological Review*, 34, 273-286
- Torgerson, W. S. (1958). *Theory and methods of scaling*. New York: Wiley.
- Thurstone, L.L. (1927). Method of paired comparison for social values. *Journal of Abnormal and Social Psychology*, 21, 384-400
- Van Zanten, V., Chiba, N., Arsmtrong, D., Barkuns, A. N., Thomson, A.B.R., Mann, V., Escobedo, S., Chakroborty, B. & Nevin, K. (2006). Validation of a 7 point global overall symptom scale to measure the severity of dyspepsia symptoms in clinical trials. *Journal Compilation*, 23, 521-529.
- Widhiarso, W. (2009). Koefisien reliabilitas pada pengukuran kepribadian yang bersifat multi dimensi. *Psikobuana*, 1, 39-48.





# Jurnal

## Penelitian dan Evaluasi Pendidikan

HIMPUNAN EVALUASI PENDIDIKAN INDONESIA (HEPI)  
kerjasama dengan

PROGRAM PASCASARJANA UNIVERSITAS NEGERI YOGYAKARTA

Alamat Redaksi : Kampus Karangmalang, Yogyakarta - 55281. Tlp. 0274 550836 psw.229

Fax : 0274 520326; e-mail: [jurnalhepi@yahoo.com](mailto:jurnalhepi@yahoo.com)