INSAN
Jurnal Psikologi dan Kesehatan Mental
http://e-journal.unair.ac.id/index.php/JPKM
p-ISSN 2528-0104 | e-ISSN 2528-5181

ARTIKEL PENELITIAN

Penggunaan *Testlet* dalam Pengembangan Tes Psikologi

WAHYU WIDHIARSO* & RETNO SUHAPTI
Fakultas Psikologi Universitas Gadjah Mada

ABSTRAK

Unit pengukuran tidak selalu berbentuk butir, akan tetapi juga dapat berbentuk kelompok butir (*testlet*). Tulisan ini mendemonstrasikan pengembangan alat ukur dengan menggunakan *testlet* yang jarang diterapkan di Indonesia. Contoh yang dipakai dalam tulisan ini adalah pengembangan pengukuran kemampuan visual bagian dari AJT Cognitive Test Battery (AJT COGTEST). Dasar pengelompokan butir menjadi satu *testlet* adalah kesamaan gambar yang diacu karena beberapa butir mengacu pada gambar yang berbeda. Teknik analisis data yang dipakai adalah model *Rasch*. Hasil perbandingan properti psikometris menunjukkan kelebihan *testlet* dibanding dengan butir. Data yang dihasilkan dari *testlet* cenderung bersifat unidimensi, tidak terjangkit dependensi lokal, memiliki ketepatan model dan daya diskriminasi butir yang lebih baik dibanding dengan unit analisis berupa butir. Hasil perbandingan fungsi informasi tes menunjukkan bahwa penggunaan *testlet* meningkatkan fungsi informasi tes. Secara umum konsep mengenai *testlet* dan aplikasinya melalui program Winsteps 3.73 dalam pengembangan alat ukur dalam dipaparkan dalam tulisan ini.

Kata kunci: *testlet, unit pengukuran, model Rasch*

ABSTRACT

The unit of analysis or measurement is not always item level, but also group of items (*testlet*). This paper demonstrates the development of measurement using *testlet* that are rarely applied in Indonesia. The example used in this paper is the development of the measurement of visual ability, one of several test included in AJT COGTEST. In this test, the basis of grouping items into *testlet* is their similarity of figure being referenced. This test consists of fifteen figure and each figure consists of seven items. The data analysis technique used is the Rasch Model. The result of comparison shows the advantages *testlet* psychometric properties as compared to item as unit of analysis. The data generated from *testlet* tends to be unidimensional, not infected local dependencies, high discrimination and high model fit than the unit of analysis in the form of grains. The comparison function test information indicates that the use *testlet* enhance test information function. In general, the concept of *testlet* and applications through Winsteps program in the development of measurement tools in presented in this paper.

Keywords: *measurement unit, Rasch model, testlet*

INSAN Jurnal Psikologi dan Kesehatan Mental, 2018, Vol. 3(1), 44-61, doi: 10.20473/jpkm.v3i12018.44-61
Dikirimkan: 30 Juli 2018 Diterima: 30 Oktober 2018 Diterbitkan: 8 November 2018
Editor: Rizqy Amelia Zein
*Alamat korespondensi: Fakultas Psikologi Universitas Gadjah Mada, Jl. Sosio-Humaniora Bulaksumur, Karang Malang Caturtunggal, Kecamatan Depok, Kabupaten Sleman Daerah Istimewa Yogyakarta 55281. Pos-el: wahyu_psy@ugm.ac.id

PENDAHULUAN

Dalam bidang ilmu sosial, sebuah sebuah tes dapat dikembangkan untuk mengukur atribut individu yang spesifik maupun yang umum. Spesifikasi tes tersebut terlihat dari seberapa luas domain perilaku yang diukur. Semakin spesifik atribut yang diukur oleh tes, semakin sempit domain perilaku diukur (DeVellis, 2011). Secara operasional spesifikasi tes dapat terlihat dari jumlah butir dan dimensi ukur yang dilibatkan dalam tes. Sebagai contoh, efikasi terhadap kesuksesan merupakan atribut yang lebih spesifik dibanding dengan kecerdasan emosi. Oleh karena itu, pengukuran efikasi hanya membutuhkan jumlah butir yang lebih sedikit dibanding dengan kecerdasan emosi. Selain itu, dari sisi dimensi yang dilibatkan, pengukuran efikasi memuat dimensi tunggal (unidimensi) sedangkan pengukuran kecerdasan emosi memuat beberapa dimensi (multidimensi).

Dari sisi format butir yang dilibatkan, pengukuran atribut yang lebih spesifik cukup menggunakan format butir yang homogen dan relatif sama sedangkan pengukuran atribut yang luas dan kompleks menggunakan format yang bervariasi. Misalnya, pengukuran efikasi dapat menggunakan format Likert untuk semua butir di dalam alat ukur. Sebaliknya pengukuran kecerdasan emosi menggunakan kombinasi antara butir yang menggunakan butir dengan format Likert dan butir dengan format tes objektif (Mayer, Salovey, Caruso, & Sitarenios, 2003). Paparan di atas menunjukkan bahwa konsekuensi penetapan atribut ukur berdampak pada panjang dan format tes. Semakin spesifik atribut yang diukur, alat ukur menjadi semakin pendek dan memiliki format yang relatif sederhana dibanding dengan yang mengukur atribut yang luas dan kompleks.

Selama ini, pengembangan pengukuran di Indonesia masih terbatas pada pengembangan pengukuran atribut yang spesifik atau secara tidak langsung 'dibuat' menjadi spesifik. Maksud dari 'dibuat' tidak spesifik adalah upaya mereduksi domain ukur yang sebelumnya ditetapkan bersifat luas menjadi bersifat spesifik. Sebagai contoh, seorang peneliti mengembangkan instrumen pengukuran etos kerja. Prosedur umum yang yang dipakai dalam melakukan analisis butir adalah mengeliminasi butir-butir yang memiliki korelasi butir-total yang rendah atau butir-butir yang dapat menurunkan reliabilitas konsistensi internal (contohnya, koefisien *Cronbach's alpha*). Prosedur seperti ini menunjukkan bahwa peneliti telah mereduksi domain ukur tes karena koefisien *Cronbach's alpha* hanya menekankan pada homogenitas butir (Boyle, 1991). Akibat pereduksian ini, validitas tes (khususnya validitas konstruk dan isi) dapat menurun karena komposisi butir dalam tes kemungkinan telah berubah dari spesifikasi awal pembuatannya. Menyederhanakan format alat ukur juga kadang dapat mereduksi validitas tes. Atribut individu seperti, inteligensi, abilitas kognitif, kemampuan bahasa atau religiusitas merupakan atribut yang kompleks sehingga membutuhkan kombinasi beberapa format. Abilitas kognitif memuat domain kemampuan verbal, logika dan numerik sedangkan kemampuan bahasa memuat penguasaan kosa kata dan struktur.

Armor (1974) mengatakan bahwa sebelum peneliti melakukan analisis butir misalnya berupa analisis daya diskriminasi dan tingkat kesukaran butir atau reliabilitas pengukuran, pengembang tes perlu melakukan analisis faktor untuk mengidentifikasi dimensionalitas pengukuran. Pernyataan ini tidak hanya berlaku pada analisis butir berbasis teori klasik, akan tetapi juga berlaku pada analisis butir berbasis teori respons butir (*item response theory*/IRT). Uji dimensionalitas pengukuran ini dapat dilakukan dengan menggunakan analisis faktor (Hair, Anderson, Tatham, & Black, 1995). Ketika analisis faktor menghasilkan beberapa dimensi, maka analisis butir dilakukan pada butir-butir dengan dimensi yang sama. Masalah unidimensionalitas pengukuran dalam pengukuran di bidang ilmu sosial telah menjadi bahasan di kalangan ahli psikometri. Berbeda dengan atribut ukur dalam bidang ilmu eksakta yang sangat spesifik, misalnya panjang dan massa benda, atribut ukur dalam bidang ilmu sosial adalah karakteristik manusia yang kompleks. Oleh karena itu, pengukuran dalam bidang ilmu sosial cenderung secara alami kompleks dan bersifat multidimensi (Hall, Snell, & Foust, 1999; Widhiarso, 2009). Meskipun pada awalnya sebuah tes disusun untuk mengukur atribut tunggal namun hasil analisis sering kali menghasilkan munculnya sub-atribut baru yang terlihat dari hasil analisis faktor yang menghasilkan dimensi yang lebih dari satu.

Untuk menghindari adanya pereduksian ini, salah satu cara yang dapat ditempuh adalah melibatkan klaster-klaster butir yang dibuat sejak awal penyusunan. Dengan kata lain, keberadaan klaster tersebut disadari sejak awal oleh peneliti dan bukan hasil eksplorasi dari data yang didapatkan. Wainer dan Wang (2000) merekomendasikan kepada para peneliti untuk melibatkan kelompok butir yang didasarkan pada stimulus yang sama pada tes yang mengukur atribut individu yang lebih kompleks. Stimulus itu berupa gambar atau cerita mengenai situasi tertentu sedangkan kelompok butir tersebut merujuk pada stimulus tersebut. Dengan menggunakan kelompok butir maka unit pengukuran telah berubah dari level butir menjadi level kelompok butir. Prosedur penggunaan unit pengukuran berupa kelompok butir telah banyak diaplikasikan di luar negeri akan tetapi belum banyak dilakukan oleh peneliti di Indonesia (Lee, Brennan, & Frisbie, 2000; Wainer & Wang, 2000). Salah satu faktor pendorongnya adalah minimnya referensi yang dapat diakses dan sedikitnya contoh pengaplikasian metode ini secara praktis. Tulisan ini bertujuan untuk mendemonstrasikan prosedur analisis dengan menggunakan kelompok butir sebagai unit pengukuran dalam pengembangan alat ukur.

Testlet Sebagai Unit Pengukuran

Berdasarkan teknik analisisnya, penggunaan unit analisis berupa kelompok butir dapat dibedakan menjadi dua jenis, Pertama, teknik analisis dengan menggunakan teori klasik. Terminologi yang sering dipakai untuk teknik ini adalah paket butir (*item parcels*) (Hall, dkk., 1999) atau kelompok butir (*item bundles*) (Rosenbaum, 1988). Kedua, teknik analisis dengan menggunakan teori respons butir yang menggunakan istilah *testlet* (Wainer & Kiely, 1987). Contoh dari pemaketan butir adalah menjumlahkan skor butir-butir yang menjadi bagian dari dimensi yang sama sedangkan contoh dari *testlet* adalah menjumlah skor butir-butir dari stimulus yang sama, misalnya berupa gambar atau skenario cerita.

Literatur menjelaskan kumpulan butir dengan terminologi yang berbeda-beda. Beberapa nama yang dimuat adalah latihan interpretatif (*interpretive exercises*), skenario (*scenarios*), *vignettes*, kelompok butir (*item bundles*), kelompok masalah (*problem sets*), *super items*, pemaketan (*parcels*) and *testlets* (Haladyna, 2004). Kumpulan butir ini biasanya dibentuk dari seperangkat butir yang mencerminkan spesifikasi isi keseluruhan untuk tes (Luecht, 2005). Gagasan awal penggunaan kelompok butir ini adalah bagaimana mengembangkan sebuah pengukuran yang sederhana akan tetapi mampu menjangkau domain atribut individu yang lebih luas. Jargon yang banyak dipakai oleh peneliti yang mengembangkan metode ini adalah "*small tests, small enough to manipulate, but big enough to carry their own context*" (Lee, dkk., 2000). Pada tulisan ini terminologi yang dipilih adalah *testlet* karena

banyak dipakai dalam pemodelan IRT yang sekaligus dipakai untuk menganalisis data yang penulis dapatkan.

Konsep *testlet* diperkenalkan oleh Wainer dan Kiely (1987) yang mendefinisikan *testlet* sebagai sekelompok butir yang terkait dengan konten tunggal yang kemudian dikembangkan menjadi sebuah satu unit analisis. Dalam pemodelan teori respons butir, kumpulan butir tersebut dinamakan dengan *testlet*, sedangkan dalam konteks pengembangan alat ukur dinamakan dengan skala mini (*mini scale*). Ditambahkan bahwa seperangkat butir tersebut biasanya merupakan sebuah alur (*predetermined paths*) yang harus diikuti oleh individu yang sedang dites. Misalnya, rentetan soal mengikuti satu stimulus tertentu baik berupa cerita atau gambar yang sama, soal-soal tersebut kemudian dapat disederhanakan menjadi satu unit yang dinamakan dengan *testlet*.

Definisi mengenai *testlet* yang dinyatakan oleh Lee, dkk. (2000) juga tidak berbeda. Mereka mendefinisikan *testlet* sebagai bagian dari butir (*subset of items*) dalam tes yang diperlakukan sebagai unit pengukuran dalam konteks konstruksi, administrasi, dan atau penyekoran tes. Dalam konteks konstruksi tes, *testlet* merupakan upaya untuk menjabarkan domain ukur menjadi konten atau komponen yang akan diukur. Pengukuran terhadap konten dari domain ukur tersebut kemudian diimplementasikan ke dalam format tertentu (misalnya skenario, *vignette* dsb.) yang dikembangkan oleh penyusun tes. Dari contoh ini dapat diketahui bahwa sebuah *testlet* juga dapat diartikan kesatuan dari seperangkat soal yang mengukur satu konten tertentu.

Dalam konteks administrasi tes, *testlet* dikaitkan dengan bagaimana tes disajikan, misalnya isu mengenai urutan soal. Proses pengurutan ini terkadang menyulitkan karena tiap soal memiliki karakter dan tingkat kesulitan yang berbeda-beda. Seperti yang telah dipaparkan di awal, soal-soal di dalam satu *testlet* telah memiliki urutan yang ditetapkan oleh penyusun. Oleh karena itu permasalahan mengenai penempatan soal dalam tes menjadi lebih mudah diatasi. Paparan ini menunjukkan bahwa *testlet* merupakan kesatuan seperangkat butir yang memiliki urutan yang telah ditetapkan. Dalam konteks penyekoran, *testlet* merupakan skor yang merupakan representasi dari kumpulan beberapa butir. Salah satu isu yang banyak dibahas pada konteks ini adalah bagaimana merepresentasikan skor kumpulan butir tersebut, apakah sebagai skor dikotomi atau politomi (Lee, Kolen, Frisbie, & Ankenmann, 2001).

Pada awalnya Wainer dan Kiely (1987) mengusulkan penggunaan *testlet* sebagai unit konstruksi dan analisis (*units of construction and analysis*) untuk tes adaptif terkomputerisasi (*computerized adaptive tests*/CATS). Perkembangan terbaru menunjukkan bahwa *testlet* sekarang dipandang sebagai solusi untuk keperluan umum untuk masalah dependensi lokal (Yen, 1993). Wainer dan Kiely (1987) memperkenalkan *testlet* sebagai pendekatan baru untuk mengatasi masalah seperti efek konteks, urutan penyajian butir, dan keseimbangan konten, yang timbul dalam pengembangan CAT. Semua masalah tersebut mengarah pada kecenderungan munculnya fenomena yang dinamakan dengan dependensi lokal.

Independensi lokal

Terkait dengan independensi lokal, di satu sisi ada penulis yang menyatakan bahwa *testlet* dapat mengatasi independensi lokal (Wainer, Bradlow, & Wang, 2007). Independensi lokal menunjukkan bahwa keberhasilan seorang partisipan untuk mengatasi sebuah soal tidak dipengaruhi keberhasilannya dalam mengatasi soal lain. Independensi lokal terjadi ketika keberhasilan subjek untuk mengatasi sebuah soal lebih dikarenakan oleh abilitas daripada faktor-faktor lain. Oleh karena itu secara statistik independensi lokal ditunjukkan dengan rendahnya korelasi antar residu yang dihasilkan dari pemodelan. Korelasi yang tinggi antar residu menunjukkan bahwa selain faktor abilitas yang

dimodelkan, masih ada faktor-faktor lain yang menyebabkan keberhasilan individu dalam mengatasi sebuah soal.

Definisi ini menunjukkan bahwa sebuah kelompok butir dengan stimulus yang sama cenderung tidak memenuhi independensi lokal. Oleh karena itu mengelompokkan rentetan butir tersebut menjadi satu unit analisis berupa *testlet* dapat mengatasi independensi lokal (Thissen & Wainer, 2001). Namun demikian beberapa ahli lainnya (Sireci, Thissen, & Wainer, 1991; Tuerlinckx & De Boeck, 2001) mengatakan bahwa *testlet* tidak sepenuhnya dapat mengatasi permasalahan independensi lokal yang dapat menyebabkan hasil estimasi yang tidak akurat dan berlebihan (*overestimate*). Ada juga yang mengatakan bahwa secara natural, butir yang secara alami berbentuk *testlet* sering tidak bersifat independen lokal ketika ditetapkan sebagai satu unit pengukuran (Thissen, Steinberg, & Mooney, 1989). Namun demikian, sifat independen lokal dalam *testlet* jarang dieksplisitkan karena sifat independen lokal sebenarnya adalah sebuah asumsi, yang merupakan sebuah kebenaran tanpa perlu diuji (Soken, 2016). Oleh karena *testlet* bersifat dependen lokal dapat secara langsung dianggap seperti dapat *testlet* yang bersifat lokal independen (Lee, dkk., 2000).

Meski independensi lokal merupakan sebuah asumsi, beberapa pendekatan telah diusulkan untuk mengatasi masalah independensi lokal pada *testlet*. Salah satu pendekatan yang paling banyak direkomendasikan adalah menentukan *testlet* sebagai satu unit pengukusran dan menetapkannya sebagai skor yang berbentuk politomi (*polytomous*) (Wainer & Lewis, 1990). Skor politomi adalah skor yang terdiri dari lebih dari dua kategori. Jika skor yang berbentuk pilah atau dikotomi terdiri dari dua kategori skor (contoh: 0 dan 1), maka skor politomi terdiri dari minimal tiga kategori (contoh: 0, 1 dan 2). Misalnya satu skenario terdiri dari tiga butir, maka skor minimal yang bisa dicapai adalah 0 (3 butir×0 = 0), sedangkan skor maksimal yang bisa dipakai adalah 3 (3 butir ×1=3).

Lebih jauh Zenisky, Hambleton dan Sireci (2000) mengatakan bahwa ketika dependensi lokal antar butir (*item dependencies*) tidak muncul pada butir-butir di dalam *testlet*, maka penetapan skor *testlet* sebagai skor politomi tidak memiliki banyak manfaat. Bahkan pada tataran tertentu akan merugikan karena fungsi informasi psikometris pengukuran menjadi berkurang seiring dengan berkurangnya jumlah parameter yang diestimasi. Sebagai contoh, sebuah tes berisi 4 *testlet* yang masing-masing berisi 2 butir dengan skor butir adalah 0 dan 1. Dengan menggunakan model IRT skor dikotomi (misalnya model 3PL), maka jumlah parameter yang diestimasi adalah 4 *testlet* × 2 butir × 3 parameter = 24 parameter. Dengan menggunakan model IRT skor politomi (misalnya model GRM), maka jumlah parameter yang diestimasi adalah 4 *testlet* × 3 parameter = 12 parameter. 3 parameter di sini didapatkan dari penjumlahan 2 ambang butir (*threshold*) dan 1 daya diskriminasi butir. Oleh karena itu, mereka menyarankan agar peneliti mendeteksi munculnya dependensi lokal antar butir dalam satu *testlet* sebelum memutuskan untuk menggunakan skor dikotomi ataukah politomi.

Model-Model Testlet

Jika dalam pemodelan IRT untuk data dikotomi peneliti dapat menggunakan model *Rasch*, 1PL, 2PL atau 3PL (Embretson & Reise, 2000) maka untuk data politomi peneliti dapat menggunakan Model Nominal (Bock, 1997), Model Respons Bergradasi (GRM) (Samejima, 1969), Model Kredit Parsial (PCM) (Masters, 1982) atau Model Generalisasi Kredit Parsial (GPCM) (Muraki, 1992). Jika respons subjek diasumsikan sebagai kategori yang tidak berurutan (misalnya, butir yang berisi empat opsi respons minat individu) maka Model Nominal lebih tepat dipakai. Sebaliknya, jika respons subjek diasumsikan merupakan urutan dari level rendah hingga tinggi, maka tiga model yang selain Model Nominal dapat diterapkan (Thissen & Wainer, 2001). Model-model yang telah dipaparkan di muka telah banyak dipakai penelitian. Untuk memilih model mana yang paling tepat peneliti dapat menguji kesesuaian data yang didapat dengan model yang ada. Tidak ada kekhususan dalam untuk tes tertentu harus menggunakan model

tertentu pula. Misalnya model GPCM dipakai untuk menganalisis *testlet* pada tes Bahasa Inggris (Y. Li, Li, & Wang, 2010), sementara itu peneliti lainnya menggunakan Model Nominal untuk tes yang serupa (Wainer, Sireci, & Thissen, 1991).

Prosedur Penyusunan Testlet

Berbeda dengan pemaketan butir (*item parceling*) yang banyak disusun dengan menggunakan nilai rerata skor dari seperangkat butir, *testlet* lebih banyak disusun berdasarkan total skor (Thissen & Wainer, 2001). Hal ini dikarenakan pemaketan butir banyak dipakai dalam pemodelan persamaan struktural (SEM) yang menekankan pada matriks kovarians sedangkan *testlet* dipakai dalam IRT yang menekankan pada pola respons subjek. Sebagai implikasi menekankan pada pola respons, maka data yang dianalisis diharapkan berbentuk bilangan bulat. Sementara itu dari sisi butir apa saja yang dikelompokkan, *testlet* juga berbeda dengan paket butir. Jika paket butir disusun dari butir-butir yang mencerminkan faktor atau dimensi yang sama, maka *testlet* disusun dari butir-butir yang memiliki stimulus yang sama, misalnya berupa skenario cerita atau gambar (Haladyna, 2004).

Lee, dkk. (2000) mengatakan bahwa jumlah butir di dalam *testlet* tidak harus seimbang, karena *testlet* lebih menekankan pada unit pengukuran dibanding dengan pengelompokan butir. Oleh karena itu sebuah tes yang berisi 20 butir dapat saja disederhanakan menjadi 10 *testlet* dengan rincian sebagai berikut: 2 *testlet* berisi 5 butir, 2 *testlet* berisi 3 butir dan 4 *testlet* berisi 1 butir. Dari contoh ini terlihat bahwa sebuah *testlet* dapat berisi satu butir. Namun demikian beberapa ahli kurang merekomendasikan penggunaan satu butir dalam satu *testlet*.

Kelebihan Testlet

Tujuan awal dari pengelompokan butir adalah menyederhanakan proses analisis, namun beberapa penelitian menunjukkan beberapa kelebihan dari pengelompokan butir. Secara umum, kelebihan utama dari menggunakan kumpulan butir adalah; (a) data yang didapatkan cenderung untuk memenuhi asumsi normalitas dibandingkan dengan data hasil dari butir tunggal. Asumsi normalitas yang dipenuhi mendukung peneliti untuk menggunakan metode estimasi kebolehjadian maksimal (*maximum likelihood*). (b) Menghasilkan nilai reliabilitas pengukuran yang tinggi dan memiliki korelasi yang tinggi dengan variabel lain (Kishton & Widaman, 1994). (c) Unit yang dianalisis menjadi lebih sederhana sehingga tidak memerlukan proses komputasi yang berat. Kesederhanaan ini terlihat dari jumlah parameter yang diestimasi dengan menggunakan unit analisis berupa *testlet* lebih sedikit dibanding dengan unit analisis berupa butir tunggal. Namun demikian, berkurangnya jumlah parameter yang diestimasi ini terkadang dapat menurunkan informasi psikometris pengukuran (Zenisky, dkk., 2000).

Ditambahkan bahwa menetapkan skor *testlet* sebagai data politomi lebih mengurangi jumlah parameter yang diestimasi dibanding dengan menetapkannya sebagai data dikotomi. Secara khusus, kelebihan *testlet* adalah kapabilitasnya untuk mengatasi dependensi lokal (lawan dari independensi lokal). Kelebihan ini berdampak pada munculnya kelebihan-kelebihan lain. Misalnya, dengan diatasinya masalah dependensi lokal maka pengukuran yang dilakukan akan lebih menghasilkan estimasi reliabilitas yang lebih akurat dan eror standar skor tes untuk tiap subjek (Wainer, dkk., 2007).

Penggunaan Testlet

Pada awalnya *testlet* banyak dipakai untuk pengukuran kemampuan bahasa karena soal cerita yang didalamnya mengandung beberapa butir (Y. Li, dkk., 2010), namun kemudian penggunaannya berkembang pada banyak bidang. Panter dan Reeve (2002) misalnya, menggunakan *testlet* dalam

pengukuran sikap terhadap rokok. Mereka menggabungkan beberapa butir yang memiliki konten serupa menjadi satu *testlet*. Misalnya butir "*Apakah Anda percaya merokok dapat membantu orang menjadi rileks?*" dan butir "*Apakah Anda percaya merokok dapat membantu mengurangi stres?*". Mereka menetapkan untuk respons subjek sebagai data kategorikal sehingga model yang memungkinkan dipakai adalah Model Kategori Nominal IRT (Bock, 1997). Pada model dikotomi, semua jenis respons subjek ditetapkan sebagai kategori eksklusif atau tidak berurutan, seperti halnya variabel lokasi tempat tinggal dalam data demografi. Oleh karena itu, dalam model ini ada empat kategori: subjek yang merespons "Tidak" pada kedua butir, subjek yang memilih "Ya" pada butir pertama, subjek yang memilih "Ya" pada butir kedua, dan subjek yang memilih "Ya" pada butir pertama dan kedua. Jika kita menetapkan keempat kategori tersebut sebagai skor politomi, maka kombinasi respons pada kedua butir tersebut dilihat sebagai data yang bersifat kontinu atau interval yang bergerak dari 0 hingga 2. Dari sini terlihat bahwa jumlah kategori menjadi berkurang dari empat kategori menjadi tiga kategori.

Testlet banyak dipakai pengembangan alat ukur untuk mengukur berbagai atribut dalam berbagai bidang. Misalnya kompetensi dalam bidang pendidikan (Hartig & Höhler, 2009), masalah karir dalam bidang industri dan organisasi (Dickinson & Tokar, 2004). Di sisi lain, penggunaan *testlet* sebagai unit analisis juga banyak dilakukan oleh peneliti. Dalam konteks ini, *testlet* tidak dikembangkan sejak awal penyusunan alat ukur akan tetapi dikembangkan dari data yang telah ada. Tujuannya adalah untuk mengatasi masalah yang terkait dengan data, misalnya ukuran sampel yang kecil atau ketidaknormalan data. Beberapa alat ukur yang telah dipakai antara lain pengukuran bagian tubuh yang terbakar (*Total Body Surface Area Burned*/TBSA), pengukuran depresi (*Beck Depression Inventory*/BDI) (Thombs, 2007) dan pengukuran kecerdasan emosi (*Mayer-Salovey-Caruso Emotional Intelligence Test*/MSCEIT) (Maul, 2011).

Tujuan Penelitian

Tulisan ini mendemonstrasikan pengembangan alat ukur dengan menggunakan *testlet*. Data yang dipakai dalam tulisan ini adalah data uji coba pengembangan tes kemampuan visual yang merupakan salah satu tes dari AJT COGTEST. AJT COGTEST yang merupakan baterai tes yang mengukur inteligensi anak-remaja berdasarkan teori Cattel Horn Carol (CHC) yang dikembangkan oleh Fakultas Psikologi UGM yang bekerja sama dengan Yayasan Dharma Bermakna (YDB) berdasarkan populasi anak-remaja Indonesia.

METODE

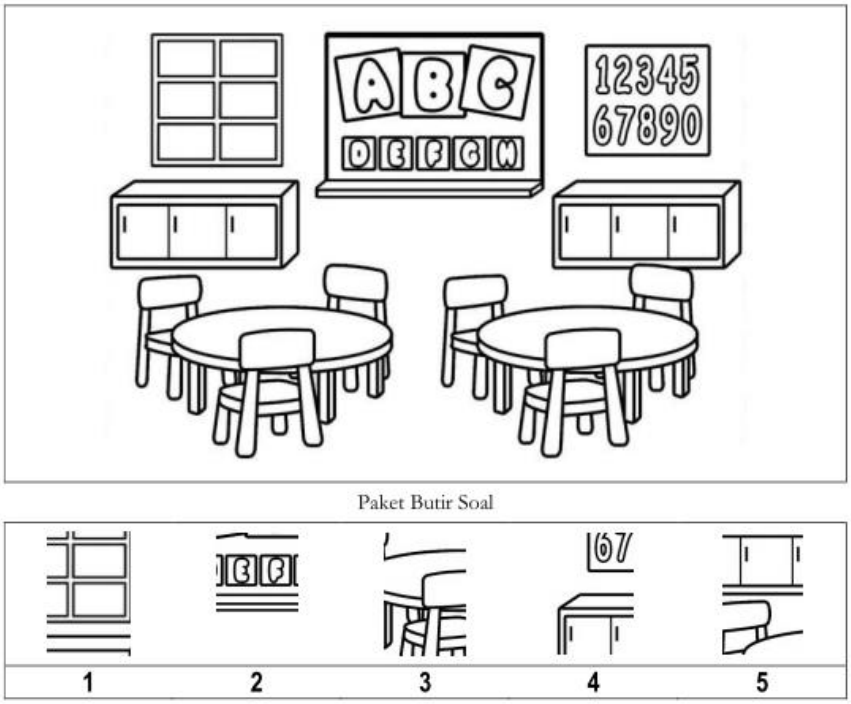
Desain Penelitian

Penelitian ini dilakukan dengan menggunakan survei model potong lintang (*cross sectional*) dengan mempertimbangkan beberapa variabel demografi berupa usia, lokasi tempat tinggal (desa/kota) dan jenis sekolah (negeri/swasta). Pengambilan data dilakukan dengan menggunakan teknik pengambilan data secara purposif berdasarkan karakteristik sampel dan keseimbangannya untuk setiap sel kategori. Data yang diambil untuk penelitian ini adalah data dari salah satu subtes dari baterai tes dari AJT Cogtest. Pengukuran dengan menggunakan AJT Cogtest dilakukan secara individual baik di kelas maupun di rumah. Persetujuan keikutsertaan dalam penelitian ini dilakukan oleh orang tua dan guru partisipan untuk anak yang masih bersekolah di tingkat dasar (SD) dan mengisi secara mandiri untuk partisipan yang sedang bersekolah mulai dari SMP hingga PT. Persetujuan tersebut diwujudkan dalam bentuk penandatanganan lembar (*inform consent*) yang telah disediakan. Waktu pengadministrasian berlangsung antara 1 jam hingga 2 jam. Beberapa pengadministrasian alat ukur dilakukan dalam waktu dua hari karena permintaan partisipan untuk menghentikan kegiatan. Penghentian kegiatan

administrasi pengukuran dilakukan di antara jeda pengukuran antara satu subtes dengan subtes lainnya.

Partisipan

Partisipan penelitian ini adalah anak-remaja yang tinggal di Daerah Istimewa Yogyakarta (DIY). Berdasarkan jenis kelaminnya, partisipan penelitian (N = 200) terdiri dari 74 orang laki-laki (M_{usia}= 11,97 SD_{usia}= 4,72) dan 126 perempuan (M_{usia}= 11,65 SD_{usia}= 4,84) dengan usia yang bergerak antara 4 hingga 20 tahun Oleh karena pemilihan partisipan dilakukan dengan mempertimbangkan keterwakilan semua rentang usia, maka distribusi pendidikan partisipan pada saat pengambilan data bersifat proporsional dengan tingkat pendidikan yang bergerak antara TK (N = 22), SD (N = 77), SMP (N = 29), SMA/SMK (N = 29) hingga perguruan tinggi (N = 27). Partisipan yang dilibatkan dalam studi kali ini adalah sebagian dari partisipan yang dilibatkan dalam proses uji coba AJT COGTEST (Purwono & Widhiarso, inpress).



Gambar 1. Contoh butir pengukuran fleksibilitas amatan

Pengukuran

Tes yang dipakai adalah tes yang mengukur abilitas visual yang merupakan salah satu dari delapan abilitas luas yang diukur oleh AJT COGTEST. Pada AJT COGTEST, abilitas visual diukur oleh empat tes yaitu kecepatan amatan (*closure speed*), ingatan visual (*visual memory*), visualisasi (*visualization*) dan fleksibilitas amatan (*flexibility of closure*). Data yang dipakai dalam penelitian ini adalah hasil pengukuran oleh tes fleksibilitas amatan. Pada tes ini partisipan diperlihatkan 15 gambar secara berurutan. Pada tiap gambar, mereka diminta untuk mencari dimana posisi potongan gambar yang ditunjukkan pada gambar yang utuh. Dengan demikian, status potongan gambar yang ditanyakan

tersebut adalah butir tes. Contoh format butir yang mengukur fleksibilitas amatan yang dapat dilihat pada Gambar 1. Butir tersebut bukan merupakan bagian dari butir AJT Cogtest akan tetapi butir yang dibuat oleh penulis untuk memberikan ilustrasi mengenai bentuk butir. Gambar tersebut memuat dua bagian, yaitu gambar utuh dan lima buah potongan gambar dari gambar utuh. Partisipan diminta untuk menunjukkan masing-masing lokasi dari lima potongan gambar di dalam gambar utuh. Setiap jawaban benar akan mendapatkan skor 1 dan jawaban salah akan mendapatkan skor 0. Dalam contoh soal tersebut, skor minimal untuk satu *teslet* tersebut adalah 0 dan skor maksimal yang dapat dicapai adalah 5. Dalam pengukuran fleksibilitas pengamatan melalui AJT Cogtest, setiap gambar memuat 15 teslet, masing-masing teslet memuat tujuh butir soal. Dengan menggunakan model penyekoran tersebut maka skor minimal untuk 15 teslet adalah 0 dan skor maksimal yang dapat dicapai adalah 105. Dasar penggunaan skor hasil pengukuran tes ini menjadi objek penelitian adalah sifat tes ini yang melanggar konsep independensi lokal; ketika siswa dapat menjawab satu butir soal maka hal ini akan memudahkan mereka untuk menjawab butir soal lainnya pada gambar tertentu. Penelitian ini mendemonstrasikan bahwa penggunaan *testlet* sebagai unit analisis akan dapat mengatasi isu ini dibanding dengan menggunakan unit analisis berupa butir soal.

Analisis Data

Analisis data dilakukan dengan menggunakan pemodelan IRT. Langkah-langkah yang dilakukan adalah:

(a) *Verifikasi Asumsi*. Memastikan beberapa asumsi penggunaan pemodelan *Rasch*/IRT terpenuhi, yaitu unidimensionalitas data dan independensi lokal. Uji dimensionalitas dilakukan dengan menggunakan analisis komponen prinsipal (PCA) pada model satu dimensi yang berisi beberapa *testlet* yang memiliki domain yang sama. Berbeda dengan analisis faktor yang dipakai pada umumnya, analisis berbasis Rasch memiliki prosedur tersendiri dalam mengidentifikasi dimensi di dalam data yang mewakili pandangan yang dianut oleh teori ini (Wright & Masters, 1982). Analisis yang dilakukan oleh Rasch tidak bertujuan untuk mengidentifikasi adanya dimensi-dimensi di dalam data akibat varians bersama antar indikator. Sebaliknya, tujuan analisis yang dilakukan adalah untuk membuktikan apakah Model Rasch yang diterapkan di dalam data dapat mendominasi varians di dalam data yang memuat varians butir dan orang. Dominasi tersebut terlihat dari kecilnya porsi dan tidak signifikannya residu varians yang dihasilkan. Kecilnya residu inilah yang menjadi fokus dalam analisis dimensionalitas dalam pendekatan Model Rasch. Uji PCA dilakukan dengan menggunakan prosedur sesuai dengan yang dijalankan oleh Winsteps 3.73. Program ini akan mengeluarkan informasi berupa proporsi varians yang dijelaskan oleh pengukuran yang dilengkapi dengan perbandingan nilai *eigenvalue* antar jumlah dimensi. Data menunjukkan merupakan hasil pengukuran unidimensi ketika nilai varians yang dijelaskan oleh pengukuran lebih dari 50% dan selisih (kontras) antar jumlah dimensi kurang dari dua (Linacre, 2000).

Uji independensi lokal dilakukan dengan mempelajari matriks korelasi antar residu butir. Korelasi yang berada di atas 0.30 menunjukkan adanya pelanggaran terhadap asumsi independensi lokal. Program yang dipakai untuk tujuan ini adalah Winsteps. Beberapa ahli seperti Embretson dan Reise (2000) mengatakan bahwa asumsi independensi lokal terkait dengan unidimensionalitas data. Jadi, ketika pengukuran bersifat unidimensi (hanya mengukur satu atribut) secara tidak langsung asumsi independensi lokal juga terpenuhi. Namun demikian, agar hasil estimasi yang dihasilkan benar-benar akurat, penulis tetap melakukan uji independensi lokal.

(b) *Indeks Ketepatan Model*. Indeks ketepatan model menunjukkan seberapa jauh performasi satu butir tes dengan model teoritik yang dianggap ideal, dalam hal ini adalah model *Rasch*. Dalam perspektif Rasch, ketepatan model ditunjukkan dengan nilai *infit* dan *outfit*. Nilai *infit* atau *outfit* yang berada di antara rentang 0.50 hingga 1.50 dianggap memiliki ketepatan model yang memuaskan.

(c) *Identifikasi Fungsi Informasi Alat Ukur*. Fungsi informasi alat ukur menunjukkan seberapa efektif alat ukur yang diberikan dalam mengukur kemampuan individu. Operasionalisasi dari kata efektif dalam hal ini adalah seberapa besar informasi mengenai abilitas individu yang dapat digali. Dalam pengembangan alat ukur berbasis IRT, efektivitas sebuah alat ukur tergantung pada level abilitas individu yang diukur. Ada alat ukur yang efektif dalam mengukur individu pada level menengah akan tetapi ada juga yang efektif dalam mengukur individu pada level tinggi. Pada tulisan ini efektivitas dua jenis analisis, yaitu analisis yang menggunakan butir dan yang menggunakan *testlet* sebagai unit pengukuran akan dibandingkan.

HASIL PENELITIAN

Hasil Analisis dengan Level Butir

Dimensionalitas. Pada program Winsteps, hasil analisis dimensionalitas data dapat dilihat melalui *Table* 23.0 yang dapat dikeluarkan dengan cara memilih *Output Tables* – *Request Subtables* – lalu menuliskan 23.0 kemudian program akan menampilkan informasi seperti yang ditampilkan pada Gambar 2.

```
     Table of STANDARDIZED RESIDUAL variance (in Eigenvalue units)
                                            -- Empirical --    Modeled
Total raw variance in observations     =        142.8 100.0%          100.0%
  Raw variance explained by measures   =         50.8  35.6%           35.9%
    Raw variance explained by persons  =         24.8  17.4%           17.5%
    Raw Variance explained by items    =         26.0  18.2%           18.3%
  Raw unexplained variance (total)     =         92.0  64.4% 100.0%    64.1%
    Unexplned variance in 1st contrast =          3.9   2.7%   4.2%
    Unexplned variance in 2nd contrast =          3.2   2.2%   3.4%
    Unexplned variance in 3rd contrast =          3.0   2.1%   3.3%
    Unexplned variance in 4th contrast =          3.0   2.1%   3.3%
    Unexplned variance in 5th contrast =          2.8   1.9%   3.0%
```

Gambar 2. Dimensionalitas Pengukuran dengan Unit Analisis Butir

Hasil analisis melalui PCA menunjukkan bahwa 105 butir yang mengukur abilitas visual cenderung bersifat multidimensi hal ini dikarenakan nilai kontras (*unexplained variance contrast*) pada perbandingan pertama hingga kedua lebih dari 2 *eigenvalue* (Linacre, 2000). Hal ini menunjukkan bahwa ada lebih dari 2 butir yang lebih mengukur dimensi selain dimensi yang diukur oleh tes. Di sisi lain variasi skor yang dapat dijelaskan oleh butir-butir dalam tes (*raw variance explained by measures*) kurang dari 50% (35.6%) yang menunjukkan eksistensi dimensi lain yang tidak diakomodasi dalam oleh skor tes cukup besar. Kesimpulan yang didapatkan adalah analisis dengan menggunakan butir sebagai unit analisis menghasilkan informasi pengukuran yang rendah karena butir-butir di dalam tes cenderung bersifat multidimensional (lihat Tabel 1).

Tabel 1. Perbandingan Hasil Analisis antara Model Analisis Berbasis Butir dan *Testlet*

Properti Psikometris	Unit Analisis Butir	Unit Analisis *Testlet*
Dimensionalitas	Nilai kontras PCA pertama, lebih dari 2 (3.90) artinya pengukuran cenderung multidimensi	Nilai kontras PCA pertama, kurang dari 2 (1.8) artinya pengukuran cenderung unidimensi
Lokal Independensi	Beberapa butir memiliki korelasi antar residu yang tinggi (>0.30) menunjukkan data tidak	Beberapa butir memiliki korelasi antar residu yang tinggi (>0.30) menunjukkan data tidak

Properti Psikometris	Unit Analisis Butir	Unit Analisis *Testlet*
	mengalami masalah terkait independensi lokal	mengalami masalah terkait independensi lokal
Daya Diskriminasi	Korelasi antara butir-total bergerak antara -0.08 hingga 0.62. 30 butir memiliki korelasi di bawah 0.3 yang menunjukkan adanya butir yang memiliki diskriminasi yang rendah	Korelasi antara butir-total bergerak antara 0.28 hingga 0.83 yang menunjukkan distribusi daya diskriminasi yang tinggi
Ketepatan Butir-Model	Ada 14 butir yang tidak memenuhi ketepatan model yang baik. Di sisi lain, ada 13 butir yang tidak dapat menghasilkan properti psikometris karena tidak memilik variasi skor.	Ada 1 butir yang kurang memenuhi ketepatan model yang baik jika ditinjau dari satu koefisien (*outfit* < 1.5) namun tidak dengan koefisien yang lain (*infit* >1.5).
Profil Fungsi Informasi	Profil fungsi informasi cukup baik karena menyebar dari individu yang memiliki level rendah hingga tinggi.	Sama seperti analisis dengan unit butir, profil fungsi informasi cukup baik namun sedikit lebih optimal.

Independensi Lokal. Pada program Winsteps, informasi mengenai independensi lokal dapat diketahui melalui *Table* 23.99 yang hasil keluarannya tercantum dalam Gambar 3. Hasil analisis melalui pengamatan terhadap korelasi antar residu terstandarisasi menunjukkan bahwa korelasi terbesar antar butir adalah 0.92. Korelasi ini ditunjukkan antara butir nomor 3 dan butir nomor 5 pada gambar tes ketiga. Artinya partisipan yang menjawab benar pada butir nomor 3 akan memberikan kesempatan yang besar untuk menjawab butir nomor 5. Di bawah nilai korelasi tersebut terdapat korelasi antar butir lain yang juga cukup tinggi (>0.30) yang menunjukkan butir-butir dalam tes tidak memenuhi prasyarat independensi lokal.

```
LARGEST STANDARDIZED RESIDUAL CORRELATIONS
USED TO IDENTIFY DEPENDENT ITEM
----------------------------------
|CORREL-| ENTRY     | ENTRY      |
|  ATION|NUMBER ITE |NUMBER ITEM |
|-------+-----------+------------|
|   .92 |    17 3.3 |    19 3.5  |
|   .56 |    43 7.1 |    54 8.5  |
|   .43 |    53 8.4 |    92 14.1 |
|   .42 |    29 5.1 |    48 7.6  |
|   .40 |    59 9.3 |    89 13.5 |
----------------------------------
```

Gambar 3. Independensi Lokal dengan Unit Analisis Butir

Indeks Ketepatan Model. Pada program Winsteps, informasi mengenai independensi lokal dapat diketahui melalui *Table* 14.1. Tabel memuat nilai tingkat kesulitan butir (*measure*) nilai *infit* dan *outfit*

serta nilai korelasi biserial. Hasil analisis menunjukkan bahwa ada banyak butir yang bermasalah. Masalah pertama adalah adanya 13 butir yang tidak dapat mengukur dengan baik. Butir tersebut tidak menghasilkan variasi skor karena semua partisipan dapat menjawab dengan benar. Masalah kedua adalah adanya 14 butir yang kurang memenuhi ketepatan model yang baik karena nilai infit maupun outfit butir tersebut berada di luar rentang penerimaan. Nilai korelasi biserial bergerak antara 0.03 hingga 0.62. Gambar 4 menampilkan sebagian informasi parameter butir tersebut.

ENTRY NUMBER	NON-EXTREME SCORE	COUNT	MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD	PTBISERL-EX CORR.	PTBISERL-EX EXP.	EXACT MATCH OBS%	EXACT MATCH EXP%	ITEM
100	130	160	-.28	.22	.90	-.8	2.48	3.0	B .48	.39	83.1	82.5	15.2
37	144	163	-1.03	.27	1.21	1.2	2.20	2.2	C .20	.36	86.5	89.1	6.2
5	165	170	-3.01	.48	1.27	.8	2.11	1.4	D .03	.24	97.1	97.1	1.5
27	137	168	-.55	.23	1.28	2.0	1.95	2.1	E .28	.46	82.1	84.5	4.6
7	97	168	1.06	.18	1.25	2.9	1.84	3.7	F .27	.45	69.6	73.4	1.7
6	83	164	1.49	.19	1.45	5.0	1.78	3.8	G .18	.43	61.6	72.9	1.6
43	156	163	-2.29	.41	1.06	.3	1.57	1.0	H .18	.26	95.1	95.8	7.1

Gambar 4. Tingkat Kesulitan dan Ketepatan Model Hasil Analisis dengan Unit Analisis Butir

Hasil Analisis dengan Level Testlet

Dimensionalitas. Hasil analisis melalui PCA menunjukkan bahwa 15 *testlet* yang mengukur abilitas visual cenderung bersifat unidimensi karena nilai kontras pada perbandingan pertama hingga kedua kurang dari 2 *eigenvalue* (Linacre, 2000). Hal ini menunjukkan bahwa kurang dari 2 butir yang lebih mengukur dimensi selain dimensi yang diukur oleh tes. Variasi skor yang dapat dijelaskan oleh butir-butir dalam tes lebih dari 50% (76.3%) yang menunjukkan dominasi dimensi utama yang diukur oleh tes cukup besar. Kesimpulan yang didapatkan adalah analisis dengan menggunakan butir sebagai unit analisis menghasilkan informasi pengukuran yang optimal karena butir-butir di dalam tes cenderung bersifat unidimensional. Hasil uji unidimensionalitas dapat dilihat di Gambar 5.

Table of STANDARDIZED RESIDUAL variance (in Eigenvalue units)

	-- Empirical --		Modeled
Total raw variance in observations =	65.0	100.0%	100.0%
Raw variance explained by measures =	50.0	76.9%	75.3%
Raw variance explained by persons =	30.1	46.3%	45.3%
Raw Variance explained by items =	19.9	30.6%	29.9%
Raw unexplained variance (total) =	15.0	23.1% 100.0%	24.7%
Unexplned variance in 1st contrast =	1.8	2.8% 12.2%	
Unexplned variance in 2nd contrast =	1.5	2.3% 10.1%	
Unexplned variance in 3rd contrast =	1.4	2.2% 9.4%	
Unexplned variance in 4th contrast =	1.3	2.0% 8.7%	
Unexplned variance in 5th contrast =	1.2	1.9% 8.1%	

Gambar 5. Dimensionalitas Pengukuran dengan Unit Analisis *Testlet*

Independensi Lokal. Hasil analisis melalui pengamatan terhadap korelasi antar residu terstandarisasi menunjukkan bahwa korelasi terbesar antar butir adalah 0.24. Dengan demikian tidak ada korelasi antar nilai residu butir yang di atas 0.30 sehingga pengukuran ini memenuhi prasyarat independensi lokal. Hasil uji independensi lokal dapat dilihat di Gambar 6.

```
LARGEST STANDARDIZED RESIDUAL CORRELATIONS
USED TO IDENTIFY DEPENDENT ITEM
----------------------------------------
|CORREL-| ENTRY         | ENTRY         |
|  ATION|NUMBER ITEM    |NUMBER ITEM    |
|-------+---------------+---------------|
|   .28 |     4 S.28.4  |     5 S.28.5  |
|-------+---------------+---------------|
|  -.29 |     3 S.28.3  |    11 S.28.11 |
|  -.26 |     6 S.28.6  |    10 S.28.10 |
|  -.25 |     6 S.28.6  |    11 S.28.11 |
|  -.22 |     5 S.28.5  |    13 S.28.13 |
----------------------------------------
```

Gambar 6. Independensi Lokal dengan Unit Analisis Butir

Indeks Ketepatan Model. Hasil analisis menunjukkan bahwa ada satu *testlet* yang kurang memenuhi ketepatan model karena memiliki nilai *outfit* yang lebih tinggi dari 1.50. Nilai korelasi biserial bergerak antara 0.29 hingga 0.84. Gambar 7 menampilkan sebagian informasi parameter butir tersebut.

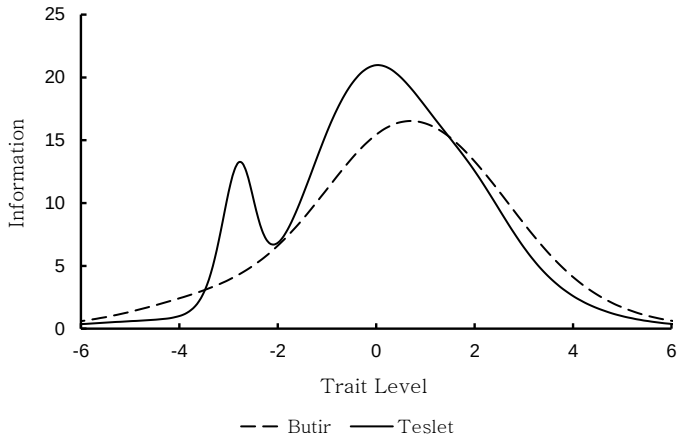
```
-------------------------------------------------------------------------------------------------
|ENTRY    NON-EXTREME           MODEL|   INFIT  |  OUTFIT  |PTBISERL-EX|EXACT MATCH|           |
|NUMBER  SCORE  COUNT  MEASURE  S.E. |MNSQ  ZSTD|MNSQ  ZSTD|CORR.  EXP.| OBS%  EXP%| ITEM    G |
|-----------------------------------+----------+----------+----------+-----------+-----------|
|     1    347    171    -.77     .12|1.60   5.4|2.23   7.3|A .29   .60| 42.1  55.0| S.28.1  0 |
|     5    897    171     .05     .07|1.08    .7|1.41   2.3|B .80   .81| 48.0  44.1| S.28.5  0 |
|     4   1102    171   -1.98     .11| .89   -.4|1.25   1.4|C .56   .58| 66.1  65.0| S.28.4  0 |
|     2    461    171   -1.92     .17|1.15    .9| .78   -.5|D .62   .62| 84.2  81.0| S.28.2  0 |
|     6    922    171    -.23     .09| .98   -.1| .96   -.3|E .79   .80| 41.5  46.1| S.28.6  0 |
|     3   1058    171   -1.82     .11| .94   -.3| .98    .0|F .62   .64| 55.6  54.6| S.28.3  0 |
```

Gambar 7. Tingkat Kesulitan dan Ketepatan Model Hasil Analisis dengan Unit Analisis Butir

Perbandingan Fungsi Informasi

Salah satu fokus pengembangan alat ukur berbasis *Rasch* atau teori respons butir adalah menyesuaikan alat ukur dengan abilitas partisipan. Jika partisipan yang diukur berada pada level menengah maka pengembangan tes diarahkan pada level abilitas tersebut. Penyesuaian ini dapat dilihat pada grafik fungsi informasi alat ukur. Grafik ini menunjukkan di level abilitas mana alat ukur akan memberikan informasi yang maksimal. Gambar 1 menunjukkan perbandingan fungsi informasi antara alat ukur sebelum dan sesudah disederhanakan menjadi *testlet*. Sumbu X pada grafik ini menunjukkan rentang level abilitas yang diukur, dari level rendah (paling kiri) hingga tinggi (paling kanan). Sumbu Y pada grafik ini menunjukkan informasi yang didapatkan. Semakin tinggi menunjukkan semakin besar informasi yang didapatkan dari hasil pengukuran. Tes yang baik adalah tes yang menghasilkan informasi yang tinggi.

Hasil analisis menunjukkan bahwa penetapan unit analisis dalam bentuk *testlet* dapat meningkatkan informasi yang didapatkan dari pengukuran. Dari bentuk grafiknya yang dapat dilihat di Gambar 8, profil fungsi informasi kedua jenis model analisis relatif setara namun grafik hasil analisis dengan menggunakan *testlet* lebih tinggi dibanding dengan butir. Pada unit analisis berupa butir, informasi maksimal didapatkan yaitu sebesar 16.54 sedangkan pada unit analisis berupa *testlet* adalah 20.98.



Gambar 8. Perbandingan Fungsi Informasi antara Butir dan *Testlet* Sebagai Unit Pengukuran

DISKUSI

Tulisan ini telah mendemonstrasikan kelebihan penggunaan *testlet* untuk tes yang tidak memenuhi asumsi independensi lokal. Data yang didemonstrasikan di tulisan ini didapatkan dari pengukuran pada tes kemampuan visual yang terdiri dari seperangkat butir yang masing-masing butir tersebut merujuk pada gambar tertentu. Tes dengan jenis ini melanggar asumsi independensi lokal, sehingga analisis butir dengan menggunakan unit analisis berupa *testlet* lebih tepat dipakai dibandingkan dengan unit analisis berupa butir. Hasil perbandingan properti psikometris menunjukkan kelebihan *testlet* dibanding dengan butir. Data yang dihasilkan dari pemaketan *testlet* cenderung bersifat unidimensi, tidak terjangkit dependensi lokal, memiliki ketepatan model dan daya diskriminasi butir yang lebih baik dibanding dengan unit analisis berupa butir.

Penggunaan *testlet* sebagai unit pengukuran berpotensi dapat meningkatkan informasi yang didapatkan oleh alat ukur. Dalam penelitian ini, penggunaan *testlet* lebih menghasilkan informasi yang lebih baik dibanding dengan butir. Temuan ini dapat dijelaskan melalui penjelasan bahwa *testlet* mengandung konten yang lebih komprehensif dibanding butir. Jika ditinjau teori klasik, pengelompokan butir (misalnya, *testlet*) tidak berbeda dengan butir karena skor tes yang dihasilkan sama. Namun hal ini menjadi berbeda ketika teknik analisis yang dipakai adalah IRT yang menekankan pada pola respons dan menggunakan model yang tidak bersifat linier. Melalui teknik IRT, dua individu (misalnya A dan B) yang memiliki skor total yang sama belum tentu memiliki kemampuan yang sama. Hal ini dapat terjadi ketika si A mampu mengatasi soal-soal yang sulit sedangkan si B hanya mengatasi soal-soal yang mudah.

Komprehensifnya menggunakan *testlet* dalam alat ukur yang dipakai dalam tulisan ini juga dikarenakan butir-butir di dalam *testlet* merupakan butir dalam rangkaian skenario yang sama. Selain memiliki kesamaan dari sisi skenario yang harus direspons, butir-butir yang dikelompokkan bersifat sekuensial. Misalnya, jawaban subjek pada butir pertama (mengenali emosi) mempengaruhi jawaban subjek pada jawaban kedua (mengenali ekspresi emosi). Selain itu, penyebab *testlet* memberikan informasi yang lebih besar juga dikarenakan sisi kedekatan waktu pengukuran. Butir-butir dalam satu *testlet* disajikan dalam satu paket sehingga memiliki waktu pengukuran yang lebih dekat. Pengelompokan butir yang memiliki kedekatan waktu pada tataran tertentu dapat mengurangi eror pengukuran (Wainer, dkk., 2007) sehingga memberikan informasi yang besar mengenai pengukuran yang dilakukan (Li, dkk., 2010).

Tingginya informasi tes dengan adanya penggunaan *testlet* juga dikarenakan *testlet* dapat mengatasi masalah independensi lokal (Wainer, dkk., 2007). Hal ini dibuktikan dengan terpenuhinya asumsi independensi lokal semua *testlet* yang dilibatkan dalam penelitian ini. Mislevy, dkk. (2012) menjelaskan bahwa dengan terpenuhinya asumsi lokal independensi maka pengukuran yang dilakukan dapat memberikan informasi yang lebih akurat. Peranan *testlet* dalam mempengaruhi informasi tes juga dinyatakan oleh beberapa ahli (misalnya, Wang, Bradlow, & Wainer, 2002). Namun demikian peranan tersebut tergantung pada alat ukur dan penyusunan *testlet*. Pada penelitian ini *testlet* meningkatkan informasi tes akan tetapi hasil ini berbeda dengan beberapa penelitian. Misalnya, Li (2010) yang mengembangkan pengukuran dengan menggunakan desain yang sama dengan penelitian ini mendapatkan bahwa penggunaan *testlet* menurunkan fungsi informasi tes. Dari sisi perubahan nilai informasi maksimal, hasil penelitian Li (2010) konsisten dengan penelitian ini. Penelitian ini menemukan bahwa dengan mengubah butir menjadi *testlet*, fungsi informasi maksimal alat ukur berubah dari level abilitas sedang menjadi menengah-rendah. Sementara itu, pada penelitian Li (2010) fungsi informasi maksimal berubah dari rendah menjadi menengah.

Penggunaan *testlet* sebagai unit analisis dapat digeneralisasikan tidak hanya dalam pengukuran kognitif saja akan tetapi juga pengukuran afektif atau kepribadian. Beberapa tes yang dikembangkan, misalnya tes pengukuran kecerdasan emosional (Mayer, dkk., 2003) juga lebih tepat dianalisis dengan menggunakan *testlet* karena di dalamnya memuat beberapa soal cerita (kasus) yang butir-butir tes mengacu pada kasus yang berbeda. Pengukuran dengan menggunakan kelompok butir yang berbeda yang mengukur domain yang lebih spesifik juga dapat menggunakan analisis berbasis *testlet*. Misalnya sebuah tes terdiri dari 12 butir yang di dalamnya memuat tiga set yang masing-masing memiliki 4 butir dengan area ukur yang berbeda juga dapat dianalisis dengan menggunakan *testlet* (Chan, Schmitt, Sacco, & DeShon, 1998). Pengukuran performasi mengajar guru yang memiliki tiga set butir (perencanaan, pelaksanaan dan evaluasi) seperti yang dilakukan oleh Hadjam dan Widhiarso (2011) juga dapat menggunakan *testlet* jika teridentifikasi terganggu oleh dependensi lokal, karena guru yang memiliki pelaksanaan pengajaran yang baik biasanya melakukan evaluasi pendidikan yang baik pula.

SIMPULAN

Pengukuran dalam bidang psikologi tidak hanya menggunakan skor soal akan tetapi juga seperangkat soal yang dinamakan dengan *testlet*. Tulisan ini menunjukkan bahwa proses analisis butir yang sesuai dengan sistem penyekoran tes akan menghasilkan keluaran analisis yang lebih tepat dibanding dengan analisis butir yang tidak sesuai sistem penyekoran tes yang ditetapkan. Administrasi butir-butir tes yang membentuk satu kelompok butir memiliki sistem penyekoran yang berbeda dengan butir-butir tes yang satu sama lainnya bersifat independen dan tidak saling mempengaruhi tanggapan atau jawaban yang diberikan oleh subjek. Dalam tulisan ini, *testlet* menjadi solusi yang tepat untuk diterapkan dalam prosedur analisis butir pada tes atau subtes yang menggunakan kelompok butir.

PUSTAKA ACUAN

Armor, D. J. (1974). Theta reliability and factor scaling. In Costner, H. L. (ed.), *Sociological Methodology*. Jossey-Bass, San Francisco. 17–50.

Bock, R. D. (1997). The nominal categories model. In W. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* New York: Springer.

Boyle, G. (1991). Does item homogeneity indicate internal consistency or item redundancy in psychometric scales? *Personality and Individual Differences, 12*(3), 291-294.

Chan, D., Schmitt, N., Sacco, J. M., & DeShon, R. P. (1998). Understanding pretest and posttest reactions to cognitive ability and personality tests. *Journal of Applied Psychology, 83*(3), 471-485. doi: 10.1037/0021-9010.83.3.471

DeVellis, R. F. (2011). *Scale development: Theory and applications*. Newbury Park: SAGE Publications, Inc.

Dickinson, J., & Tokar, D. M. (2004). Structural and discriminant validity of the career factors inventory. *Journal of Vocational Behavior, 65*(2), 239-254. doi: http://dx.doi.org/10.1016/j.jvb.2003.07.002

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists : Multivariate applications book series* Mahwah (NJ): Lawrence Erlbaum Associates, Inc.

Hadjam, M. N. R., & Widhiarso, W. (2011). Efikasi Mengajar Sebagai Mediator Peranan Faktor Kepribadian Terhadap Performansi Mengajar Guru. *Humanitas, 3*(2), 1-16.

Hair, J. F., Anderson, R. E., Tatham, R. L., & Black, W. C. (1995). *Multivariate Data Analysis*. New Jersey: Prentice Hall.

Haladyna, T. M. (2004). *Developing and validating multiple-choice test items*. Mahwah, NJ: Lawrence Erlbaum Associates.

Hall, R. J., Snell, A. F., & Foust, M. S. (1999). Item Parceling Strategies in SEM: Investigating the Subtle Effects of Unmodeled Secondary Constructs. *Organizational Research Methods, 2*(3), 233-256. doi: 10.1177/109442819923002

Hartig, J., & Höhler, J. (2009). Multidimensional IRT models for the assessment of competencies. *Studies in Educational Evaluation, 35*(2–3), 57-63. doi: http://dx.doi.org/10.1016/j.stueduc.2009.10.002

Kishton, J. M., & Widaman, K. F. (1994). Unidimensional versus domain representative parceling of questionnaire items. An empirical example. *Educational and Psychological Measurement*, 54, 757-765.

Lee, G., Brennan, R. L., & Frisbie, D. A. (2000). Incorporating the *Testlet* Concept in Test Score Analyses. *Educational Measurement: Issues and Practice, 19*(4), 9-15. doi: 10.1111/j.1745-3992.2000.tb00041.x

Li, T. (2010). *IRT Modelling situational judgment tests: Items vs. testlets* (Poster). Oxford: . Hogrefe Ltd.

Li, Y., Li, S., & Wang, L. (2010). Application of a General Polytomous *Testlet* Model to the Reading Section of a Large-Scale English Language Assessment *Research Report No. ETS RR-10-21*. Princeton, N.J: Educational Testing Service.

Linacre, J. M. (2000). *WINSTEPS, version 3.02.* Chicago: Winstep.com.

Luecht, R. M. (2005). Computer-Adaptive Testing. In B. Everitt & D. C. Howell (Eds.), Encyclopedia of Statistics in Behavioral Science. Hoboken, N.J: Wiley.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47*(2), 149-174.

Maul, A. (2011). Examining the structure of emotional intelligence at the item level: New perspectives, new conclusions. *Cognition & Emotion, 26*(3), 503-520. doi: 10.1080/02699931.2011.588690

Mayer, J. D., Salovey, P., Caruso, D. R., & Sitarenios, G. (2003). Measuring emotional intelligence with the MSCEIT V2.0. *Emotion, 3*(1), 97-105. doi: 10.1037/1528-3542.3.1.97

Mislevy, J. L., Rupp, A. A., & Harring, J. R. (2012). Detecting Local Item Dependence in Polytomous Adaptive Data. *Journal of Educational Measurement*, 49(2), 127-147. doi:10.1111/j.1745-3984.2012.00165.x

Muraki, E. (1992). A Generalized Partial Credit Model - Application of an Em Algorithm. *Applied Psychological Measurement, 16*(2), 159-176.

Panter, A. T., & Reeve, B. B. (2002). Assessing tobacco beliefs among youth using item response theory models. *Drug and Alcohol Dependence, 68, Supplement*(0), 21-39. doi: http://dx.doi.org/10.1016/S0376-8716(02)00213-2

Purwono, U. & Widhiarso, W. (inpress). Technical Manual AJT Cogtest. Jakarta: PT Melintas Cakrawala Indonesia – Fakultas Psikologi UGM

Rosenbaum, P. R. (1988). Items bundles. *Psychometrika, 53*(3), 349-359. doi: 10.1007/bf02294217

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores *Psychometrika Monographs*.

Soken, N. (2016). Creating Design Value Through Understanding Human Cognition and Behavior. *Design Management Review*. 27,12-18. doi:10.1111/drev.12003

Sireci, S. G., Thissen, D., & Wainer, H. (1991). On the Reliability of *Testlet*-Based Tests. *Journal of Educational Measurement, 28*(3), 237-247. doi: 10.1111/j.1745-3984.1991.tb00356.x

Thissen, D., Steinberg, L., & Mooney, J. A. (1989). Trace Lines for *Testlet*s: A Use of Multiple-Categorical-Response Models. *Journal of Educational Measurement, 26*(3), 247-260. doi: 10.2307/1434990

Thissen, D., & Wainer, H. (2001). *Test Scoring*. Mahwah, NJ.: Lawrence Erlbaum Associates, Inc.

Thombs, B. D. (2007). Use of the Beck Depression Inventory for assessing depression in patients hospitalized with severe burn Disentangling symptoms of depression from injury and treatment factors. *Burns, 33*(5), 547-553. doi: http://dx.doi.org/10.1016/j.burns.2006.10.398

Tuerlinckx, F., & De Boeck, P. (2001). The effects of ignoring item interactions on the estimated discrimination parameters in item response theory. *Psychological Methods, 6*, 181-195.

Wainer, H., Bradlow, E. T., & Wang, X. (2007). *Testlet response theory and its applications*. New York, NY: Cambridge University Press.

Wainer, H., & Kiely, G. L. (1987). Item Clusters and Computerized Adaptive Testing: A Case for *Testlet*s. *Journal of Educational Measurement, 24*(3), 185-201. doi: 10.2307/1434630

Wainer, H., & Lewis, C. (1990). Toward a psychometrics for *testlet*s. *Journal of Educational Measurement, 27*(1), 1-14. doi: 10.1111/j.1745-3984.1990.tb00730.x

Wainer, H., & Wang, X. (2000). Using a new statistical model for *testlet*s to score TOEFL. *Journal of Educational Measurement, 37*(3), 203-220. doi: 10.1111/j.1745-3984.2000.tb01083.x

Wang, X., Bradlow, E. T., & Wainer, H. (2002). A General Bayesian Model for *Testlet*s: Theory and Applications. *Applied Psychological Measurement, 26*(1), 109-128. doi: 10.1177/0146621602026001007

Widhiarso, W. (2009). Koefisien reliabilitas pada pengukuran kepribadian yang bersifat multidimensi. *Psikobuana, 1*(1), 39 - 48.

Wright, B. D. & Masters, G. N. (1982). *Rating Scale Analysis*. Chicago: MESA Press.

Yen, W. M. (1993). Scaling performance assessments: Strategies for Managing Local Item Dependence. *Journal of Educational Measurement, 30*(3), 187-213.

Zenisky, A. L., Hambleton, R. K., & Sireci, S. G. (2000). *Effects of Local Item Dependence on the Validity of IRT Item, Test, and Ability Statistics*. Paper presented at the Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.