ISSN 2442-3041
Math Didactic: Jurnal Pendidikan Matematika
Vol. 1, No. 3, September - Desember 2015

METODE KALIBRASI DAN DESAIN TES BERDASARKAN TEORI RESPONS BUTIR (IRT)[2]

Dina Huriaty
Dosen STKIP PGRI Banjarmasin Pendidikan Matematika
E-mail: dina_rty@yahoo.co.id

Abstrak: Ketika akan menyusun perangkat soal untuk suatu tes, pengembang dapat menggunakan butir soal yang telah dikalibrasi ditambah dengan butir-butir soal baru. Pada keadaan ini, permasalahan yang muncul adalah bagaimana menempatkan parameter butir yang baru atau parameter butir pada tes sebelumnya, apakah perlu untuk ditempatkan pada skala butir-butir yang telah dikalibrasi atau pada skala yang baru. Cara untuk menempatkan parameter estimasi dari dua kelompok yang terpisah kedalam skala yang sama, dapat dilakukan dengan menghitung parameter estimasi untuk setiap kelompok dan kemudian mengubah skala dengan menggunakan common-items. Hal ini dapat dilakukan melalui proses kalibrasi. Butir-butir yang telah dikalibrasi ditempatkan sebagai butir bersama atau common-items pada perangkat soal yang baru. Ada tiga cara kalibrasi yaitu kalibrasi terpisah (*separate calibration*), kalibrasi serentak (*concurrent calibration*), dan kalibrasi tetap (*fixed calibration*).

Kata kunci: kalibrasi, tes, teori respons butir

Kalibrasi butir adalah proses estimasi untuk menentukan parameter-parameter butir berdasarkan model *Item Response Theory* (IRT). *Item Response Theory* atau teori respons butir merupakan teori tentang bagaimana variabel orang dan variabel butir menentukan data respons ketika seseorang menjawab butir tersebut (Umar, 1999).

Teori respons butir mempunyai kelebihan dibandingkan teori tes klasik, yaitu statistik butir tidak tergantung pada kelompok, skor tes yang diperoleh dapat menggambarkan kemampuan individu, tidak memerlukan tes yang paralel untuk menghitung koefisien reliabilitas, dan dapat menyediakan ukuran yang tepat untuk setiap skor kemampuan. Teori respons butir didasarkan pada dua postulat, yaitu: (1) kemampuan setiap individu pada suatu butir soal dapat diperkirakan dengan seperangkat faktor yang disebut karakteristik laten (*latent-traits*), (2) hubungan antara kemampuan

[2] Disampaikan pada seminar internal jurusan/program studi pendidikan matematika STKIP PGRI Banjarmasin 20 Februari 2016

individu pada suatu butir dan perangkat kemampuan yang mendasarinya sesuai dengan grafik fungsi monoton, yang disebut kurva karakteristik butir (*ICC= Item Characteristic Curve*). Kurva karakteristik butir menunjukkan bahwa semakin tinggi tingkat kemampuan individu, semakin besar peluang menjawab benar suatu butir.

Pembahasan

A. Kalibrasi dan teori respon butir

Asumsi yang mendasari teori respons butir adalah unidimensi, independensi lokal, dan invarian. Asumsi unidimensi menyatakan bahwa pada setiap tes, hanya ada satu kemampuan yang diukur oleh butir-butir tes tersebut. Hal ini berarti bahwa asumsi unidimensi dipenuhi jika butir tes hanya mengukur satu kemampuan. Asumsi unidimensi pada praktiknya tidak dapat dipenuhi secara ketat karena adanya faktor-faktor kognitif, kepribadian, dan faktor pelaksanaan tes, seperti kecemasan, motivasi, dan kecenderungan untuk menebak.

Asumsi lain pada teori respons butir adalah independensi lokal. Independensi lokal terjadi jika kemampuan yang mempengaruhi performansi tes bersifat konstan, artinya respons peserta tes dalam menjawab suatu butir tes bebas secara statistik terhadap respons peserta tes itu dalam menjawab butir lainnya. Asumsi independensi lokal menyatakan bahwa tidak ada korelasi antara respons peserta tes pada butir soal yang berbeda. Hal ini menunjukkan bahwa kemampuan yang dinyatakan dalam model merupakan satu-satunya faktor yang mempengaruhi respons peserta tes terhadap butir soal. Jika faktor-faktor yang mempengaruhi prestasi konstan, maka respons subjek terhadap pasangan butir yang manapun akan independen secara statistik satu sama lain. Asumsi independensi lokal akan terpenuhi jika jawaban peserta terhadap suatu butir soal tidak mempengaruhi jawaban peserta terhadap butir soal yang lain.

Menurut Hambleton, Swaminathan, & Rogers (1991: 10), independensi lokal secara matematis dinyatakan sebagai berikut.

$$P(u_1, u_2, \dots, u_n|\theta) = P(u_1|\theta).P(u_2|\theta) \dots . P(u_n|\theta) = \prod_{i=1}^{n} P(u_i|\theta)$$

Keterangan:

$P(u_1, u_2, \dots, u_n|\theta)$: probabilitas peserta tes dengan kemampuan θ menjawab benar butir ke-*1* hingga butir ke-*n*.

$P(u_i|\theta)$: probabilitas peserta tes dengan kemampuan θ menjawab benar butir ke-*i*.

i : nomor butir tes = 1, 2, 3, ..., *n*

n : banyaknya butir tes.

Invariansi parameter adalah karakteristik butir soal yang tidak tergantung pada distribusi parameter kemampuan peserta tes dan parameter ciri peserta tidak tergantung pada ciri butir soal. Hal ini berarti bahwa kemampuan seseorang tidak akan berubah hanya karena mengerjakan tes yang berbeda tingkat kesulitannya dan parameter butir tidak akan berubah hanya karena diujikan pada kelompok peserta tes yang berbeda tingkat kemampuannya.

Invariansi parameter kemampuan dapat diketahui dengan mengadministrasikan dua perangkat tes atau lebih yang mempunyai tingkat kesulitan yang berbeda pada sekolompok peserta tes. Invariansi parameter kemampuan akan diperoleh jika hasil estimasi kemampuan peserta tes tidak berbeda pada setiap perangkat tes yang diujikan. Demikian pula halnya dengan invariansi parameter butir tidak akan berubah jika diujikan pada kelompok peserta tes yang berbeda-beda kemampuannya.

Selain memenuhi persyaratan unidimensi, independensi lokal, dan invariansi parameter, hal lain dalam teori respons butir yang juga perlu diperhatikan adalah penentuan model respons. Teori respons butir menggunakan pendekatan probabalistik untuk menyatakan hubungan antara kemampuan individu dengan harapan menjawab benar. Model distribusi yang digunakan adalah distribusi logistik.

Ada tiga model logistik dalam teori respos butir, yaitu model logistik satu, dua, dan tiga parameter. Model-model ini sesuai untuk data respons butir yang diskor dikotomus (Hambleton, Swaminathan, & Rogers, 1991: 12). Yang membedakan ketiga model ini adalah banyaknya parameter yang digunakan untuk menggambarkan karakteristik butir pada setiap model logistiknya atau parameter-parameter butir. Parameter-parameter butir tersebut adalah indeks kesukaran butir (*b*), indeks daya beda butir (*a*), dan tebakan semu (*c*). Ketiga unsur ini berhubungan sehingga menghasilkan fungsi atau lengkungan respons yang disebut juga kurva karakteristik butir.

Hubungan ini diartikan bahwa ada suatu butir yang direspon oleh sejumlah peserta tes. Respons peserta tes terhadap butir tersebut ditentukan oleh ciri butir dan ciri peserta tersebut. Ciri individu dinyatakan sebagai parameter θ dan ciri butir dinyatakan sebagai parameter butir *a*, *b*, dan *c*. Respons individu terhadap butir dinyatakan dalam bentuk probabilitas jawaban yang benar $P(\theta)$. Hubungan ini untuk butir ke-*j*, dinyatakan sebagai berikut.

$$P_j(\theta) = \varphi(\theta, a_j, b_j, c_j)$$

$P_j(\theta)$ menyatakan probabilitas jawaban benar untuk butir ke-*j*, parameter θ menyatakan ciri individu yang dapat berupa kemampuan akademik peserta tes, parameter daya beda butir, parameter kesukaran butir, dan parameter faktor tebakan semu atau *pseudo-guessing*.

Parameter ciri individu θ membentuk suatu kontinum yang membentang tidak terbatas, yaitu $-\infty \leq \theta \leq +\infty$. Hasil pengukuran di bidang pendidikan sering menyebar pada bentuk distribusi probabilitas normal, di mana nilai baku yang digunakan terletak antara -3 hingga +3. Untuk mengestimasi parameter-parameter pada teori respons butir diperlukan suatu ukuran banyaknya peserta yang cukup besar. Ukuran banyaknya peserta untuk mengestimasi parameter tergantung dari model yang dipakai. Model tiga parameter memerlukan peserta tes yang lebih besar dibandingkan jika menggunakan model dua parameter atau satu parameter. Demikian juga halnya jika menggunakan model dua parameter, maka peserta tes yang diperlukan lebih besar dibandingkan jika menggunakan model satu parameter.

Model logistik tiga parameter adalah model yang paling umum dari ketiga model. Dengan kurva berbentuk S dan asimptut yang lebih rendah, model ini sangat tepat ketika individu dengan kemampuan rendah terkadang dapat merespons dengan benar butir yang sulit (Hulin, Drasgow, & Parsons, 1983: 29). Tes-tes kemampuan yang menggunakan format pilihan ganda dan instrumen sikap adalah contoh situasi dimana model logistik tiga parameter cocok untuk digunakan. Peserta tes cenderung untuk memilih jawaban terbaik yang mereka anggap paling menarik jika mereka tidak dapat menemukan jawabannya.

Manfaat potensial penggunaan pendekatan IRT dalam analisis butir tes terutama dalam proses kalibrasi butir. Ketika akan menyusun perangkat soal untuk suatu tes, pengembang dapat menggunakan butir soal yang telah dikalibrasi ditambah dengan

butir-butir soal baru. Pada keadaan ini, permasalahan yang muncul adalah bagaimana menempatkan parameter butir yang baru atau parameter butir pada tes sebelumnya, apakah perlu untuk ditempatkan pada skala butir-butir yang telah dikalibrasi atau pada skala yang baru. Cara untuk menempatkan parameter estimasi dari dua kelompok yang terpisah kedalam skala yang sama, dapat dilakukan dengan menghitung parameter estimasi untuk setiap kelompok dan kemudian mengubah skala dengan menggunakan *common-items*. Hal ini dapat dilakukan melalui proses kalibrasi. Butir-butir yang telah dikalibrasi ditempatkan sebagai butir bersama atau *common-items* pada perangkat soal yang baru.

Menurut *Standards for Educational and Psychological Testing* (1999:172), kalibrasi dalam IRT merupakan proses estimasi parameter-parameter dari fungsi respons suatu butir. Fungsi respons suatu butir memuat dua parameter, yaitu parameter butir dan parameter orang. Menurut Wells, Subkoviak, & Serlin (2002), proses kalibrasi digunakan untuk mengestimasi parameter butir soal, dan mengamati kemampuan butir-butir soal dalam membedakan antartingkat *trait* laten. Menurut Yen & Fitzpatrick (2006: 129), kalibrasi butir adalah mengestimasi parameter, yaitu menentukan estimasi parameter butir dan parameter kemampuan data respons butir pada model IRT. Dengan demikian kalibrasi adalah proses estimasi parameter butir dan parameter orang untuk mengetahui kedudukan butir dan orang dalam suatu instrumen tes berdasarkan model IRT.

Ada tiga cara kalibrasi yaitu kalibrasi terpisah (*separate calibration),* kalibrasi serentak (*concurrent calibration*), dan kalibrasi tetap (*fixed calibration*). Pada beberapa penelitian (Li, et al , 1997; Ban, et al, 2001, Taehoon & Petersen, 2009), kalibrasi *fixed parameter* disebut sebagai *fixed item parameter calibration* dan *fixed abc*. Li, et al. (1997). menyatakan bahwa kalibrasi *Fixed ABC* menghasilkan hasil penyetaraan yang lebih stabil, terutama untuk parameter *c* dan θ pada penyetaraan horizontal. Hasil kalibrasi *fixed parameter* terhadap estimasi θ dan kalibrasi terpisah secara umum konsisten. Hasil simulasi menunjukkan bahwa kalibrasi *fixed parameter* dan kalibrasi serentak menghasilkan estimasi parameter butir dan kemampuan yang sangat akurat dan stabil. Hasil penelitian tersebut juga menunjukkan bahwa kalibrasi *fixed parameter* menghasilkan estimasi yang lebih akurat dan stabil dibandingkan kalibrasi serentak pada tes yang relatif mudah. Menurut Taehoon & Petersen (2009) kalibrasi *fixed parameter* menunjukkan hasil yang konsisten dibandingkan dua metode kalibrasi yang lainnya, yaitu kalibrasi serentak (*concurrent calibration*) dan kalibrasi terpisah (*separate calibration*).

1. Kalibrasi terpisah

Pada metode kalibrasi terpisah, parameter-parameter butir pada setiap tes diestimasi secara terpisah atau sendiri-sendiri (Hanson & Beguin, 2002). Parameter-parameter butir diestimasi secara terpisah untuk setiap kelompok dari beberapa kelompok atau kelompok-kelompok yang tidak ekivalen. Parameter-parameter yang dihasilkan pada setiap kelompok tidak pada skala umum. Untuk mendapatkan skala umum yang didasari pada satu skala yaitu skala (0,1), maka skala lainnya yang berasal dari kalibrasi terpisah harus dikonversi terlebih dahulu kedalam skala dasar.

Suatu desain tes memerlukan transformasi linear ketika menggunakan metode kalibrasi terpisah. Parameter butir pada *common items* antara dua kelompok digunakan untuk mengestimasi transformasi

skala parameter, misalnya parameter butir pada kelompok *target* ditempatkan pada skala parameter butir pada kelompok *base*.

Empat metode untuk mentransformasi skala digunakan pada model dikotomus teori respons butir. Metode tersebut adalah metode Momen, metode Kurva karakteristik, metode Kai-kuadrat minimum, dan metode Kuadrat terkecil (Kolen & Brennan, 2004). Pada metode Momen termasuk didalamnya metode Mean/Mean dan metode Mean/Sigma, sedangkan metode Kurva karakteristik termasuk didalamnya metode Karakteristik kurva dari *Haebara* dan metode karakteristik tes dari *Stocking* dan *Lord*.

Metode Mean/Sigma menggunakan nilai rata-rata hitung dan simpangan baku untuk mengestimasi parameter kesulitan butir dari *common items* pada tes *I* dan *J*. Secara matematis dinyatakan pada formulasi berikut.

$$A = \frac{\sigma(b_J)}{\sigma(b_I)}$$

$$B = \mu(b_J) - A\ (b_I)$$

Keterangan:

A dan B : koefisien penyetaraan

$\sigma(b_J)$: simpangan baku parameter *b* untuk butir-butir pada skala J

$\sigma(b_I)$: simpangan baku parameter *b* untuk butir-butir pada skala I

$\mu(b_J)$: nilai rata-rata hitung parameter *b* untuk butir-butir pada skala J

$\mu(b_I)$: nilai rata-rata hitung parameter *b* untuk butir-butir pada skala I

Metode Mean/Mean menggunakan nilai rata-rata hitung dari estimasi parameter *a* dan *b*. Formulasinya ditunjukkan pada persamaan berikut.

$$A = \frac{\mu(a_I)}{\mu(a_J)}$$

$$B = \mu(b_J) - A\ (b_I)$$

Keterangan:

$\sigma(a_I)$: nilai rata-rata hitung parameter *a* untuk butir-butir pada skala I

$\sigma(a_J)$: nilai rata-rata hitung parameter *a* untuk butir-butir pada skala J

$\mu(b_J)$: nilai rata-rata hitung parameter *b* untuk butir-butir pada skala J

$\mu(b_I)$: nilai rata-rata hitung parameter *b* untuk butir-butir pada skala I

Metode dari *Haebara* menggunakan penjumlahan kuadrat selisih antara kurva karakteristik butir untuk setiap butir pada individu dengan kemampuan θ. Jika kemampuan θ diketahui, maka jumlah kuadrat selisih semua butir secara matematis dinyatakan pada persamaan berikut.

$$H \quad (\theta_i) = \sum_j \left[p_i\ (\theta_J \mid \bar{a}_J\ , \bar{b}_J\ \hat{c}_J\ ) - p_i\ \left(\theta_J \mid \frac{\bar{a}_I}{A}, \widehat{A^{-1}}_I\ , B, \hat{c}_I\ , \right) \right]^2$$

Perbedaan antara kurva karakteristik butir pada dua skala adalah kuadrat dan jumlah dari semua butir. H adalah kumulasi semua peserta tes untuk mendapatkan nilai konstanta transformasi, A dan B, berdasarkan kriteria dapat disederhanakan menjadi persamaan berikut.

$$H \quad = \sum_i H \quad (\theta_i)$$

Stocking dan Lord (1983) menggunakan kuadrat selisih antara kurva karakteristik tes pada kemampuan θ, dinyatakan pada formulasi berikut ini.

$$S \quad (\theta_i) = \left[\sum_j p_i\ (\theta_J \mid \bar{a}_J\ , \bar{b}_J\ \hat{c}_J\ ) - \sum_J p_{iJ} \left(\theta_J \mid \frac{\bar{a}_I}{A}, \widehat{A^{-1}}_I\ , B, \hat{c}_I\ , \right) \right]^2$$

Setiap kurva karakteristik butir dari semua *common items* dapat digunakan untuk menghitung kurva karakteristik tes. Selisih antara kurva karakteristik butir pada dua skala kemudian dikuadratkan, sehingga $S \quad if$ semua peserta tes dijumlahkan untuk menemukan konstanta transformasi A dan B yang dapat disederhanakan kedalam persamaan berikut.

$$S \qquad = \sum_{i} S \qquad (\theta_i)$$

Beberapa penelitian menunjukkan bahwa transformasi dengan metode *Stocking* dan *Lord* dan metode *Haebara* menghasilkan estimasi lebih stabil dibandingkan metode Mean/Mean dan metode Mean/Sigma.

2. Kalibrasi serentak

Kalibrasi serentak mengestimasi parameter pada semua butir dan pada semua tes pada satu kali proses estimasi dan menempatkan semua estimasi parameter pada skala yang sama, yaitu (0,1) atau pada skala umum. Ketika kalibrasi serentak dilakukan, penting untuk menggunakan program estimasi yang memungkinkan dapat mengkalibrasi beberapa kelompok secara bersamaan atau serentak (Kolen & Brennan, 2004), seperti Bilog-MG dan Multilog. Kedua program ini menggunakan kebolehjadian maksimum marjinal untuk mengestimasi parameter yang cocok dengan model logistik tiga parameter. Semua data diestimasi dan hasil semua estimasi ditempatkan pada skala umum melalui satu kali proses estimasi.

3. *Fixed parameter calibration*

Metode *fixed parameter calibration* menghasilkan skala bersama dengan cara menetapkan parameter *common items* kemudian mengestimasi parameter *common items* dan butir yang bukan butir bersama untuk kemudian ditempatkan pada skala yang sama. Terdapat dua metode *fixed calibration*, yaitu metode *fixed C* dan metode *fixed ABC*. Pada *fixed C*, estimasi parameter *c* dari tes referens digunakan sebagai nilai awal untuk tes *target*, dan keduanya tidak diestimasi lagi, sedangkan parameter *a* dan *b* diestimasi. Setelah estimasi parameter butir, proses untuk menemukan nilai A dan B yang digunakan pada transformasi linear, sama seperti pada metode kalibrasi terpisah.

Prinsip dasar dari metode *fixed ABC* adalah menetapkan estimasi parameter *a*, *b*, dan *c* pada *common item*s dari tes sebelumnya dan kemudian mengestimasi parameter butir sisa yang bukan *common item*s bersama-sama dengan *common item*s, sehingga butir sisa yang bukan *common item*s berada pada skala yang sama dengan *common item*s. Metode *fixed ABC* banyak digunakan pada kalibrasi *on-line* yang digunakan pada CAT (Ban, et al., 2001) Metode ini juga digunakan pada pengembangan bank soal (Li, et al, 1997). Metode *fixed ABC* dan metode kalibrasi serentak, mengestimasi parameter berdasarkan respons kelompok peserta yang diakumulasi dari beberapa tes, sehingga ukuran sampel menjadi relatif besar. Keadaan ini meminimalisir masalah ketidakakuratan estimasi parameter *b* dan *c*, yang mungkin terjadi pada kelompok peserta tes dengan kemampuan rendah. Jadi metode *fixed ABC* mempunyai beberapa sifat yang juga dimiliki oleh metode kalibrasi serentak dan hal ini memungkinkan diperoleh hasil pengkaitan yang lebih stabil dibandingkan metode kalibrasi terpisah yang menggunakan transformasi skala.

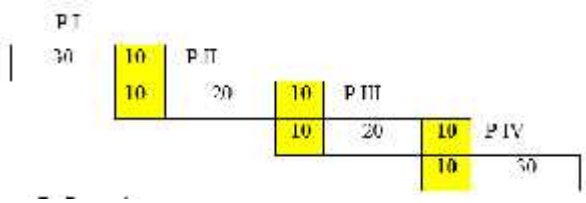
B. Desain Tes

Pada tes-tes standar, diasumsikan bahwa parameter butir telah diketahui. Hal ini dikarenakan kalibrasi butir dilakukan selama proses standarisasi tes. Pada tes yang baru

diujikan, parameter-parameter butir akan diestimasi dari data yang diperoleh. Selain mengestimasi parameter butir, parameter kemampuan juga diestimasi. Pada berbagai situasi diperlukan lebih dari satu perangkat tes yang diujikan (Wright & Stone, 1979: 98). Pada situasi ini kadangkala butir bersama (*common items*) ditempatkan pada perangkat tes tersebut. Sebagai contoh, jika ingin mengestimasi 60 butir sedangkan hanya diperlukan 40 butir yang akan diujikan pada setiap peserta, maka 60 butir dapat dibagi menjadi tiga set tes, yaitu A, B, dan C. Dua perangkat tes dapat dibentuk, perangkat pertama memuat A dan B, sedangkan perangkat kedua memuat A dan C. Kedua perangkat diujikan pada dua kelompok peserta tes yang berbeda. Respons kedua kelompok terhadap butir-butir pada tes A dapat digunakan untuk membuat skala umum terhadap parameter semua butir yang diestimasi. Sejumlah *common items* digunakan sebagai pengait untuk menghitung estimasi parameter butir dari dua buah perangkat pada skala yang sama (Lee & Ban, 2010). Penempatan skala umum pada dua atau lebih tes memungkinkan untuk membandingkan tingkat kesukaran tes dan juga menjadi dasar dalam pengembangan bank soal. Menurut Hambleton, Swaminathan, & Rogers (1991: 128) ada empat desain untuk pengkaitan yang dapat digunakan untuk penskalaan parameter butir, yaitu: (1) *Single-Group Design*, (2) *Equivalent-Group Design*, (3) *Anchor-Test Design*, dan (4) *Common-Person Design*, sedangkan menurut Kolen & Brennan (1995: 13) ada tiga desain pengumpulan data, yaitu: (1) *Random-Group Design*, (2) *Single-Group Design*, dan (3) *Common-Items Nonequivalent-Group Design*. Setiap desain memiliki kelebihan dan keterbatasan dalam pengadministrasiannya.

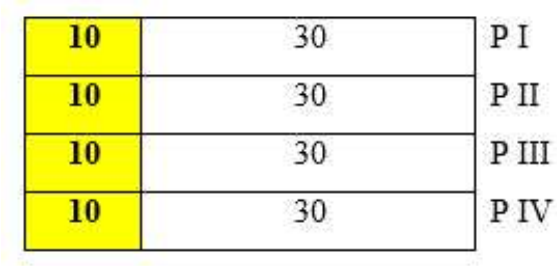
Pada *Single-Group Design*, dua buah tes yang akan dikaitkan diujikan pada kelompok yang sama. Desain ini sederhana, tetapi tidak praktis karena membutuhkan waktu yang lama dalam pengadministrasiannya. Desain yang lebih praktis dalam pengadministrasiannya dan tidak ada efek akibat mengulang dan kelelahan adalah *Equivalent-Group Design*. Pada desain ini dua buah tes yang akan dikaitkan, diujikan pada dua kelompok yang ekuivalent. Pemilihan kelompok dilakukan secara acak. Pada *Anchor-Test Design*, tes-tes dengan butir bersama diberikan kepada dua kelompok uji. Setiap tes mempunyai sejumlah butir bersama. Desain ini mudah dilaksanakan dan seringkali digunakan. Jika pemilihan *common-item* dilakukan dengan tepat maka desain ini dapat menjadi alternatif untuk mengatasi masalah pada desain *Single-Group* dan *Equivalent-Group*. Pada desain *Common-Person*, dua buah tes yang dikaitkan diberikan pada dua kelompok dengan kelompok bersama mendapatkan kedua tes tersebut. Kelemahan dari desain ini adalah akan terjadi efek kelelahan pada kelompok bersama, karena kelompok tersebut akan mendapatkan tes dalam jumlah yang lebih banyak dibandingkan peserta yang tidak termasuk dalam kelompok bersama. Letak butir bersama dalam perangkat tes pada desain ini digambarkan pada Gambar 1.

Alternatif 1.



Keterangan: P = perangkat

Alternatif 2.



10	30	P I
10	30	P II
10	30	P III
10	30	P IV

Gambar 1. Desain Tes yang Memuat *Common Items*

Pada gambar alternatif 1 tampak contoh empat perangkat tes yang digabungkan dengan 30 butir inti. Jumlah soal seluruhnya setiap perangkat adalah 40 butir. Pada perangkat I terdiri dari 10 butir inti dan 30 butir yang bukan butir inti, 10 butir inti ini adalah soal yang menghubungkan perangkat tes I dengan perangkat tes II. Perangkat tes II terdiri dari 20 butir yang bukan butir inti dan 20 butir inti, 10 butir menghubungkan perangkat tes II dengan perangkat tes I dan 10 butir menghubungkan perangkat tes II dengan perangkat tes III. Hal ini sama dengan perangkat tes III, sedangkan tes IV terdiri dari 30 butir yang bukan butir inti dan 10 butir inti yang menghubungkan perangkat tes IV dengan perangkat tes III. Soal inti yang menghubungkan setiap perangkat berbeda, sehingga jumlah seluruhnya butir inti adalah 30 butir yang berfungsi sebagai *common items*. Pada gambar alternatif 2, tampak bahwa empat perangkat tes hanya mempunyai 10 butir inti yang berfungsi *common items* dan masing-masing perangkat ditambahkan 30 butir yang bukan butir inti.

Common items mempunyai peranan penting dalam kalibrasi butir, karena itu ketika desain *anchor-test* digunakan, hendaknya memperhatikan sifat dan karakteristik dari *common items* dan penggunaan skornya. *Common items* seharusnya menggambarkan miniatur tes yang disetarakan dan item tersebut relatif berada pada nomor urut yang sama, baik pada naskah tes yang pertama maupun naskah tes lainnya. Jumlah *common items* disarankan minimal 20% dari panjang tes untuk model tes yang diskor secara dikotomus (Kolen & Brennan, 1995:248). Secara umum desain *common test* memerlukan ukuran sampel yang besar (Peterson, Kolen & Hoover, 1989). Desain *anchor-test* ini sering digunakan, karena dua kelompok peserta tes yang dibutuhkan tidak harus sama kemampuannya dan dapat berasal dari populasi yang berbeda.

Kesimpulan

Kalibrasi adalah proses estimasi parameter butir dan parameter orang untuk mengetahui kedudukan butir dan orang dalam suatu instrumen tes berdasarkan model IRT. Ada tiga cara kalibrasi yaitu kalibrasi terpisah (*separate calibration)*, kalibrasi serentak (*concurrent calibration*), dan kalibrasi tetap (*fixed calibration*).

Pengembang tes dapat menggunakan berbagai alternatif desain tes. Penempatan parameter estimasi dari dua kelompok yang terpisah kedalam skala yang sama, dapat dilakukan dengan cara menghitung parameter estimasi untuk setiap kelompok dan kemudian mengubah skala menggunakan *common-items*. Cara yang lain adalah dengan menetapkan parameter *common-items* kemudian mengestimasi parameter butir yang bukan *common-items* secara bersama-sama, sehingga berada pada skala yang sama dengan *common-items*.

Kalibrasi dan desain tes digunakan dalam pengembangan bank soal dan pengembangan tes *on-line* seperti *CAT (Computerized Adaptive Testing)*.

Daftar Pustaka

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.

Ban, J-C., Hanson, B.A., Tianyou Wang, et al. (2001) A comparative study of on-line pretest item-calibration/scaling methods in computerized adaptive testing. *Journal of Educational Measurement*, 38, 191-212.

Hambleton, R. K., Swaminathan, H. & Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park, CA: Sage Publications, Inc.

Hanson, B.A. & Beguin, A.A. (2002). Obtaining a common scale for item response theory item parameters using separate versus concurrent estimation in the common-item equating design. *Applied Psychological Measurement*, 26, 3-24.

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

Kolen, M. J. & Brennan, R. L. (1995). *Test equating: Methods and practices.* New York: Springer.

Kolen, M. J. & Brennan, R. L. (2004). *Test equating, scaling, and linking: Methods and practices (2nd ed.).* New York: Springer.

Lee, W-C & Ban, J-C. (2010). A comparison of IRT linking procedures. *Applied Measurement in Education*, 23, 23-48.

Li,-Y. H., Griffith, W. D., & Tam, H.P. (1997, June). *Equating multiple tests via an IRT linking design: Utililizing a single set of common items with fixed common item parameters during the calibration process.* Paper presented at the annual meeting of the psychometric society, Knoxville, TN.

Petersen, N.S., Kolen, M.J., & Hoover, H.D. (1989). Scaling, norming, and equating. Dalam Robert. L. Linn (Ed.) *Educational Measurement. 3rd ed.* (pp. 221-262). Washington, DC: American Council on Education.

Taehoon Kang & Petersen, N. (2009). Linking item parameters to a base scale. *ACT Research Report Series*, 2009-2. Diambil tanggal 20 September 2010, dari http://www.act.org/ research/researchers/reports/pdf/ACT_ RR2009-2.pdf.

Umar, J. (1999). Item banking. Dalam G.N. Masters & J.P. Keeves (Eds.). *Advances in measurement in educational research and assessment* (pp. 207-218) Oxford: Elsevier Science Ltd.

Wells, C.S., Subkoviak, M.J., & Serlin,K.C. (2002). The effect of item parameter drift on examinee ability estimates. *Applied Psychological Measurement*, 26, 77-87.

Wright, B. D. & Stone, M. H. (1979). *Best test design.* Chicago: Mesa Press.