Perbandingan Validitas Kriteria *Test of English Proficiency* terhadap *ITP-TOEFL* Antara Korelasi Biasa dengan Korelasi Kanonis

Heri Retnawati (retnawati_heriuny@yahoo.co.id)
Pendidikan Matematika FMIPA Universitas Negeri Yogyakarta

Abstrak

Salah satu kriteria pengembangan tes standar adalah validitas tes, baik berdasarkan isi, konstrak, maupun kriteria. Untuk mengetahui validitas kriteria dari suatu tes, skor tes peserta yang telah menempuh tes ini dikorelasikan dengan skor tes peserta ketika menempuh tes yang lebih standar. Biasanya korelasi yang digunakan yaitu korelasi biasa, yang mengorelasikan skor tes yang dikembangkan dengan skor tes yang lebih standar. Pada realitasnya, tes yang dikembangkan terdiri dari subtes-subtes tertentu, demikian pula tes kriterianya. Alternatif metode analisis yang dapat digunakan yaitu metode korelasi kanonis untuk mengetahui validitas kriteria subtes-subtes yang dikembangkan dengan subtes-subtes kriteria. Pada tulisan ini akan disajikan perbandingan korelasi biasa dan korelasi kanonis untuk mengetahui validitas kriteria *Test of English Proficiency* (TOEP) buatan Indonesia terhadap ITP-TOEFL. Hasil analisis menunjukkan bahwa validitas kriteria yang diketahui dengan korelasi biasa cenderung lebih ..... dibandingkan dengan korelasi kanonis, yang menunjukkan bahwa .....

Kata kunci: validitas kriteria, korelasi, korelasi kanonis, *TOEP, TOEFL*

Pendahuuan

Validitas merupakan hal yang penting dalam menentukan kualitas tes. Ada berbagai pendapat mengenai validitas. Menurut *American Educational Research Association, American Psychological Association, and National Council on Measurement* in Education (AERA, APA, and NCME) dalam *Standards for Educational and Psychological Testing*, validitas merujuk pada derajat dari fakta dan teori yang mendukung interpretasi skor tes, dan merupakan pertimbangan paling penting dalam pengembangan tes (1999). Ahli lain mengemukakan bahwa validitas suatu alat ukur adalah sejauhmana alat ukur itu mampu mengukur apa yang seharusnya diukur (Nunnally, 1978, Allen & Yen, 1979: 97; Kerlinger, 1986; Syaifudin Azwar, 2000: 45). Sementara itu, Linn & Gronlund (1995) menjelaskan validitas mengacu pada kecukupan dan kelayakan interpretasi yang dibuat dari penilaian, berkenaan dengan penggunaan khusus. Pendapat ini diperkuat oleh Messick (1989) bahwa validitas merupakan kebijakan evaluatif yang terintegrasi tentang sejauhmana fakta empiris dan alasan teoretis mendukung kecukupan dan kesesuaian

inferensi dan tindakan berdasarkan skor tes. Berdasarkan beberapa pendapat tersebut, dapat disimpulkan bahwa validitas akan menunjukkan dukungan fakta empiris dan alasan teoretis terhadap terhadap interpretasi skor tes, dan terkait dengan kecermatan pengukuran.

Validitas itu dapat dikelompokkan menjadi tiga tipe, yaitu: (1) validitas kriteria (*criterion-related*), (2) validitas isi, dan (3) validitas konstruk (Nunnally, 1978, Allen & Yen, 1979, Fernandes, 1984, Woolfolk & McCane, 1984, Kerlinger, 1986, dan Lawrence, 1994). Validitas ini dapat diketahui melalui fakta keberadaaan validitas. Sumber fakta validitas dapat dikelompokkan menjadi isi tes, proses respons, struktur internal, hubungan dengan variabel lain, dan konsekuensi dari pelaksanaan tes (AERA, APA, and NCME, 1999; Cizek, et al., 2008). Keberadaan validitas dari suatu perangkat tes ini dapat diketahui melalui analisis isi tes dan analisis empiris dari skor tes data respons butir (Lissitz & Samuelsen, 2007).

Validitas isi suatu instrumen adalah sejauhmana butir-butir dalam instrumen itu mewakili komponen-komponen dalam keseluruhan kawasan isi objek yang hendak diukur dan sejauh mana butir-butir itu mencerminkan ciri perilaku yang hendak diukur (Nunnally, 1978; Fernandes, 1984). Sementara itu Lawrence (1994) menjelaskan bahwa validitas isi itu keterwakilan pertanyaan terhadap kemampuan khusus yang harus diukur. Berdasarkan hal ini, dapat disimpulkan bahwa validitas isi terkait dengan analisis rasional terhadap domain yang hendak diukur untuk mengetahui keterwakilan instrumen dengan kemampuan yang hendak diukur.

Validitas konstruk adalah validitas yang menunjukkan sejauhmana instrumen mengungkap suatu kemampuan atau konstruk teoretis tertentu yang hendak diukurnya (Nunnally, 1978, Fernandes, 1984). Prosedur validasi konstruk diawali dari suatu identifikasi dan batasan mengenai variabel yang hendak diukur dan dinyatakan dalam bentuk konstruk logis berdasarkan teori mengenai variabel tersebut. Dari teori ini ditarik suatu konskuensi praktis mengenai hasil pengukuran pada kondisi tertentu, dan konskuensi inilah yang akan diuji. Apabila hasilnya sesuai dengan harapan maka instrumen itu dianggap memiliki validitas konstruk yang baik.

Pada tes prestasi belajar dan tes kompetensi, validitas merupakan syarat yang sangat diperlukan dalam pengembangan tes. Menurut pendapat Sireci yang didukung Lissitz & Samuelsen (2007), validasi tes yang dipergunakan dalam dunia pendidikan sebaiknya melibatkan analisis isi tes dan analisis empiris dari skor tes dan data respons butir. Analisis isi tes terkait dengan validitas isi yang selanjutnya diperlukan

juga analisis empiris untuk mengetahui validitas konstruk. Kedua analisis ini dimaksudkan agar tes di dunia pendidikan memenuhi syarat tes yang standar.

Validitas berdasarkan kriteria dibedakan menjadi dua, yaitu validitas prediktif dan validitas konkuren. Fernandes (1984) mengatakan validitas berdasarkan kriteria dimaksudkan untuk menjawab pertanyaan sejauh mana tes memprediksi kemampuan peserta di masa mendatang *(predictive validity)* atau mengestimasi kemampuan dengan alat ukur lain dengan tenggang waktu yang hampir bersamaan *(concurrent validity)*. Hal senada juga disampaikan oleh Lawrence (1994) yang mengatakan bahwa tes dikatakan memiliki validitas prediktif bila tes itu mampu memprediksikan kemampuan yang akan datang. Dalam analisis validitas prediktif, performansi yang hendak diprediksikan disebut dengan kriteria. Besar kecilnya harga estimasi validitas prediktif suatu instrumen digambarkan dengan koefisien korelasi antara prediktor dengan kriteria tersebut.

Validitas kriteria diketahui dengan mengestimasi korelasi skor tes peserta dengan skor kriteria. Korelasi ini disebut dengan koefisien validitas (Linn & Gronlund, 1995), yang menyatakan derajat hubungan antara prediktor dengan kriteria.

Korelasi biasa....

Korelasi kanonis

Salah satu manfaat dengan adanya validitas kriteria yakni dapat memprediksikan suatu skor kemampuan ke skor kriteria dalam rangka memprediksikan kemampuan atau performen peserta tes. Prediksi ini dilakukan melalui persamaan regresi.

Ada dua macam regresi yang dapat digunakan. Model yang pertama yakni regresi sederhana atau regresi tunggal, dengan prediktor hanya satu variabel saja (Pedhazur, 1973, Kleinbaum, dkk.,1988; Walpole, dkk., 2002). Model ini dituliskan dengan

$$\hat{Y} = b_0 + b_1 X \quad \text{..... (1)}$$

dengan $\hat{Y}$ merupakan hasil prediksi, b_0 konstanta, b_1 koefisien prediktor, dan X merupakan prediktor.

Model yang kedua yakni regresi ganda, dengan prediktor lebih dari satu variabel. Pada kasus kedua ini, digunakan jika tes terdiri dari beberapa subtes, dan prediktor merupakan jumlahan skor dari subtes-subtes yang berada dalam seperangkat tes. Model regresi ganda dengan dua prediktor disajikan pada persamaan 2.

$$\hat{Y} = b_0 + b_1 X_1 + b_2 X_2 \quad \text{.....(2)}$$

dengan $\hat{Y}$ merupakan hasil prediksi, b_0 konstanta, b_1 koefisien prediktor pertama, X_1 prediktor pertama, b_2 koefisien prediktor kedua, dan X_2 merupakan prediktor kedua. Kedua model ini belum dibandingkan yang paling akurat, untuk memprediksikan skor kriteria kemampuan peserta tes.

Tantangan dunia global yang sarat dengan muatan persaingan mengisyaratkan bahwa seseorang yang ingin berhasil dalam mengarungi dunia nyata perlu memiliki kemahiran yang diakui oleh dunia global. Terkait dengan hal ini, prestasi seseorang juga mesti diukur dengan cara dan hasil yang dapat diakui oleh dunia global. Mengikuti alur berpikir ini, kemahiran berbahasa Inggris siswa SMA juga perlu diukur dengan cara dan hasil yang diakui tidak hanya di Indonesia tetapi di mancanegara juga sehingga mereka akan memiliki kesempatan untuk melanjutkan studi dan/atau mencari kerja tidak hanya di negeri sendiri tetapi juga di mancanegara juga.

Selama ini alat ukur yang digunakan untuk mendapatkan informasi tentang kemahiran berbahasa Inggris adalah tes bahasa Inggris yang dibuat oleh lembaga asing, misalnya TOEFL, TOEIC, dan IELTS. Biaya untuk mengikuti tes ini cukup mahal, tetapi memang hasilnya jelas diakui di semua Negara karena memang tes-tes tsb bersifat standar, yang telah dikembangkan melalui serentetan kegiatan yang ditujukan untuk menjaga agar tes yang dihasilkan memenuhi kriteria tes yang baik. Jika setiap siswa di Indonesia diharapkan mengikuti tes, sebagian besar dari mereka tidak akan mampu untuk membiayainya. Jika Negara yang dibebani biaya tes tsb, jelas kurang pas karena pada dasarnya Negara bukan penanggung biaya kegiatan seperti itu. Maka satu hal yang merupakan kekuarangan/kelemahan yang menjadi kendala untuk meminta siswa mengikuti tes bahasa Inggris standar internasional

adalah masalah kekurangan biaya. Kelemahan lain adalah ketergantungan dunia pendidikan pada pihak asing. Hal ini berdampak buruk pada pembentukan kepribadian Indonesia yang kokoh.

Untuk mengatasi kedua kelemahan tersebut di atas, Dit PSMA memandang perlu untuk segera mengembangkan tes profisiensi bahasa Inggris (*Test of English Proficiency* atau TOEP), yang mengukur kemahiran menggunakan bahasa Inggris dalam dunia nyata para lulusan SMA. Pada tahun 2007 telah dimulai pengembangan seperangkat instrumen pengukuran kemahiran menggunakan bahasa Inggris tersebut, yang dilanjutkan tahun 2008 dan 2009. Selama 3 tahun (2007-2009) telah dikembangkan 7 perangkat TOEP yang diberi nama TOEP 1, 2A, 2B, 3A, 3B, 4, dan 5 yang saling paralel.

Tes Kemahiran Bahasa Inggris (*Test of English Proficiency, TOEP*) yang merupakan tes standar untuk mengukur kemahiran berbahasa Inggris siswa Sekolah Menengah Atas (SMA). TOEP yang dikembangkan merupakan tes tertulis (*paper and pencil test*) pada tahun 2007 dan 2008, dan selanjutnya dirintis tes untuk mengukur kemampuan *Speaking* dan *Reading* di tahun 2009 dan 2010. Penskoran tiap butir dilakukan dengan sistem dikotomi, benar diberi skor 1 dan jika salah diberi skor 0. Tes ini khusus mengukur kemahiran siswa SMA dalam menggunakan bahasa Inggris, khususnya *Reading* dan *Listening*. Tes terdiri dari 100 butir soal, dengan rincian 50 butir tes *Reading* dan 50 butir tes *Listening*. Terkait dengan tes yang dikembangkan merupakan tes standar internasional, pada kegiatan ini juga dihasilkan petunjuk pelaksanaan TOEP. Hal ini dimaksudkan agar setiap TOEP yang dilaksanakan benar-benar merupakan tes yang terstandar.

TOEP dikembangkan melalui proses menjabarkan tujuan menjadi indikator-indikator, yang kemudian dikembangkan menjadi butir. Ini berarti TOEP memenuhi syarat tes yang baik ditinjau dari validitas isinya. Validitas kenampakan (*face validity*) untuk menjadi tes yang baik juga terpenuhi, mengingat pengembangan tes ini mulai dari menyusun butir sampai dengan perakitan tes melibatkan ahli yang terkait, baik dari perguruan tinggi maupun dari praktisi di lapangan (guru). Validitas lain yang digunakan yakni validitas *criterion-related evidence of validity* jenis konkuren, yakni mengaitkan skor TOEP dengan skor TOEFL Institusional perolehan siswa.

Terkait dengan adanya validitas *criterion-related evidence of validity* jenis konkuren yang dimiliki TOEP, skor perolehan siswa SMA yang menempuh TOEP dapat dikonversikan ke skor tes lain, misalnya TOEFL. Hasil konversi ini dapat

dimanfaatkan siswa untuk keperluan pendaftaran/seleksi masuk ke perguruan tinggi dalam negeri atau ke dunia kerja. Mengingat TOEP ini bersertifikasi internasional, sertifikat yang diperoleh siswa juga akan sangat bermanfaat bagi peserta didik bila akan melanjutkan ke Perguruan Tinggi atau memasuki dunia kerja yang memerlukan kemahiran berbahasa Inggris di luar negeri.

Sehubungan dengan adanya dua jenis korelasi yang dapat digunakan untuk mengetahui validitas kriteria, yaitu korelasi biasa dan korelasi kanonis, maka pada tulisan ini akan dibandingkan hasil kedua jenis analisis ini untuk mengetahui validitas kriteria kemampuan bahasa Inggris siswa SMA terhadap TOEFL. Pada korelasi biasa dikorelasikan skor TOEP terhadap skor TOEFL, dan pada korelasi kanonis dikorelasikan skor *Listening* dan *Reading* dari TOEP untuk dikorelasikan dengan skor *Listening* dan *Reading* dari TOEFL.

Tujuan

Pada tulisan ini dibahas perbandingan hasil korelasi biasa dan korelasi kanonis untuk mengetahui validitas kriteria tes kemampuan bahasa Inggris siswa SMA (TOEP) dengan kriteria ITP-TOEFL.

Metode

Untuk mengetahui validitas kriteria TOEP dengan kriteria TOEFL, dilakukan dengan pendekatan kuantitatif. Data yang digunakan merupakan data dokumetasi skor TOEP dan skor TOEFL pada 833 siswa SMA di Indonesia yang menempuh kedua tes tersebut. Distribusi peserta disajikan pada Tabel 1.

Tes TOEFL dilakukan tidak terlalu jauh jarak waktu pelaksanaannya dengan tes TOEP. Setelah keduanya diskor, kemudian dibuat diagram pencar untuk memprediksi adanya korelasi skor TOEP dengan skor TOEFL dan keberadaan hubungan linear antara keduanya. Selanjutnya diestimasi korelasi biasa antara skor TOEP dengan skor TOEFL, dan korelasi kanonis antara skor Listening dan skor Reading TOEP dengan skor Listening dan skor Reading pada skor TOEFL.

Tabel 1. Distribusi Peserta untuk *Benchmarking* TOEP dengan TOEFL

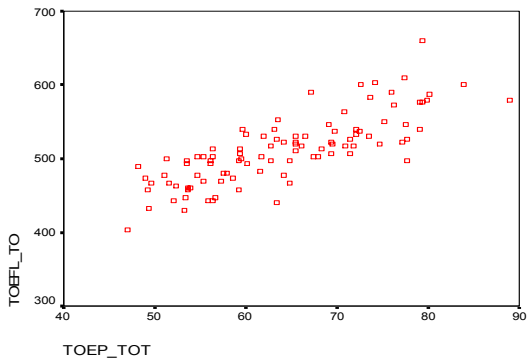
Pelaksanaan	Perangkat TOEP	Banyaknya Peserta (N)	Peserta

Feb 2008	1	98	Yogyakarta, Jawa Timur, Jawa Barat, DKI Jakarta
Desember 2008	2A	145	Banten, Lampung, Yogyakarta, Jawa Timur, Bali, Sulawesi Utara
	2B	150	Sumatera Barat, Riau Kepulauan, Jawa Barat, Jawa Tengah, Kalimantan Timur, Papua
	3A	115	Banten, Lampun, Yogyakarta, Jawa Timur, Bali, Sulawesi Utara
	3B	139	Sumatera Barat, Riau Kepulauan, Jawa Barat, Jawa Tengah, Kalimantan Timur, Papua
November 2009	4	78	Riau, Sulawesi Tengah
	5	108	Bangka Belitung, DKI Jakarta, Kalimantan Selatan

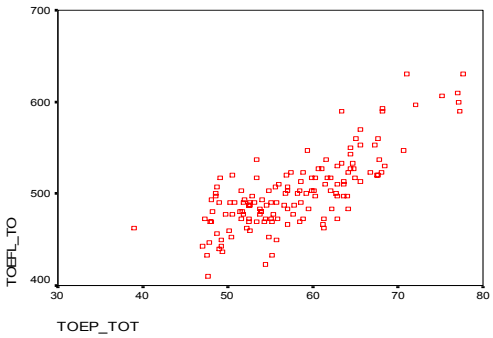
Hasil

Untuk mengetahui keberadaan hubungan linear antara variabel prediktor dengan variabel kriteria, dibuat diagram pencar (*Scatter Plot*) terlebih dahulu. Pada model regresi tunggal, variabel prediktornya merupakan skor TOEP dan variabel kriterianya merupakan skor TOEFL. Hasilnya disajikan pada Gambar 1. Demikian pula pada model regresi ganda, variabel prediktornya merupakan skor TOEP *Listening* dan skor TOEP *Reading* dan variabel kriterianya merupakan skor TOEFL, dengan hasil pada Gambar 2.

TOEP 1

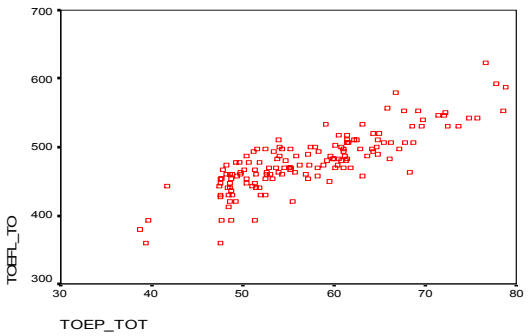


TOEP 2A



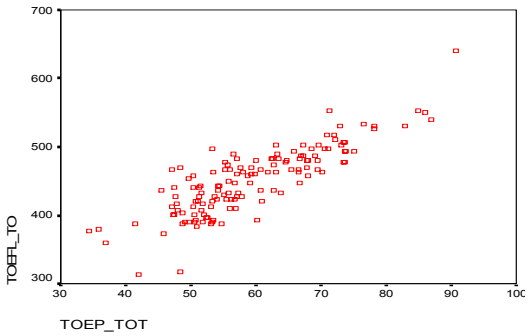
TOEP 2B

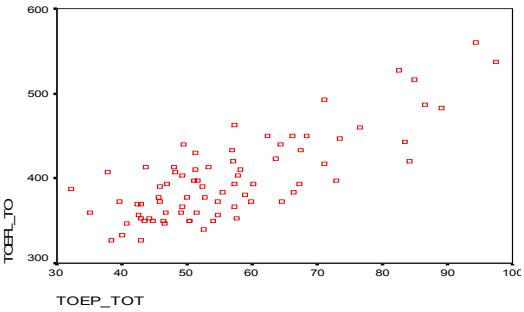
TOEP 3A



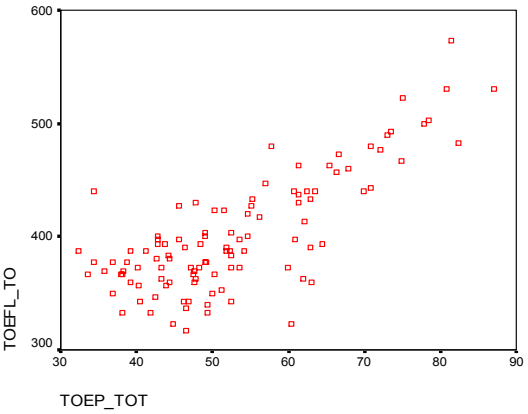
TOEP 3B

TOEP 4



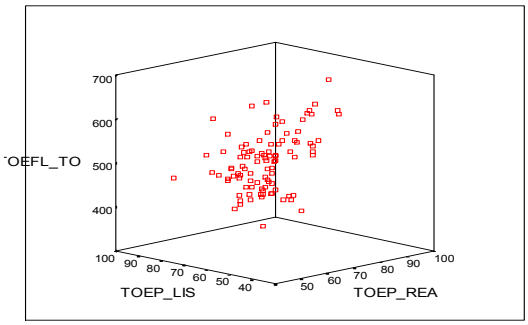


TOEP 5

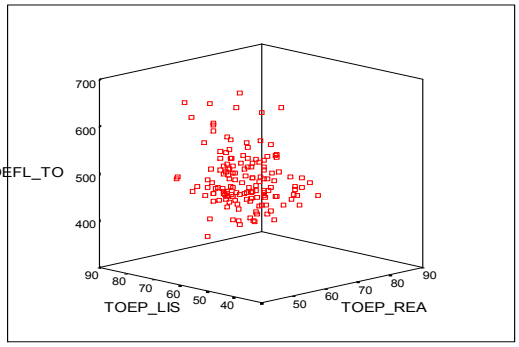


Gambar 1. Diagram Pencar Skor TOEP untuk Memprediksi Skor TOEFL

TOEP 1

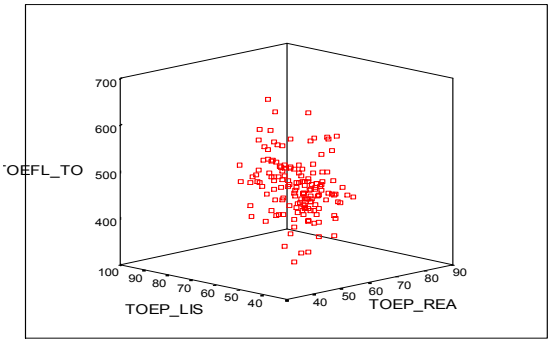


TOEP 2A

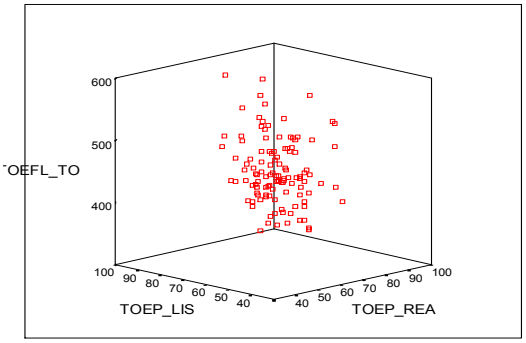


TOEP 2B

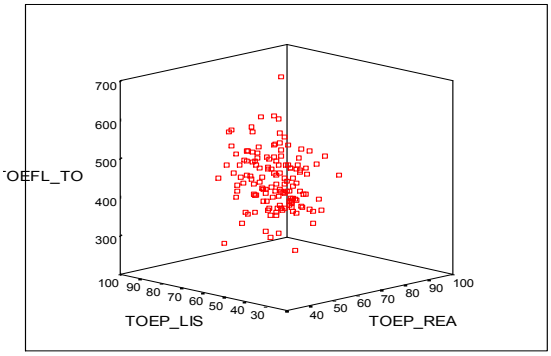
TOEP 3A

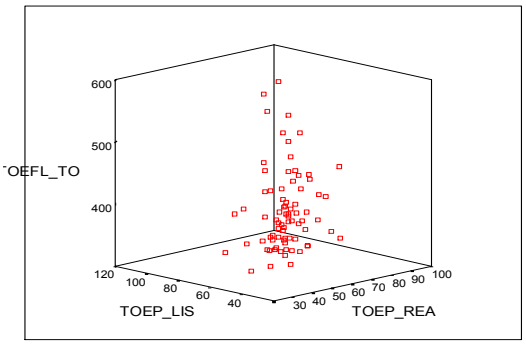


TOEP 3B

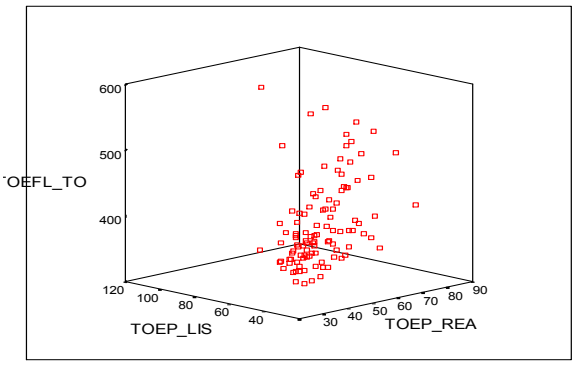


TOEP 4





TOEP 5



Gambar 2. Diagram Pencar TOEP *Listening* dan TOEP *Reading* untuk Memprediksi Skor TOEFL

Mencermati diagram pencar pada Gambar 1, diperoleh bahwa terdapat hubungan linear antara skor TOEP dengan skor TOEFL pada model regresi tunggal. Pada Gambar 2 juga menunjukkan adanya hubungan linear antara skor TOEP *Listening* dan TOEP *Reading* untuk Memprediksi skor TOEFL. Hasil estimasi korelasi baik pada model regresi tunggal maupun regrasi ganda disajikan pada Tabel 2 dan Gambar 3.

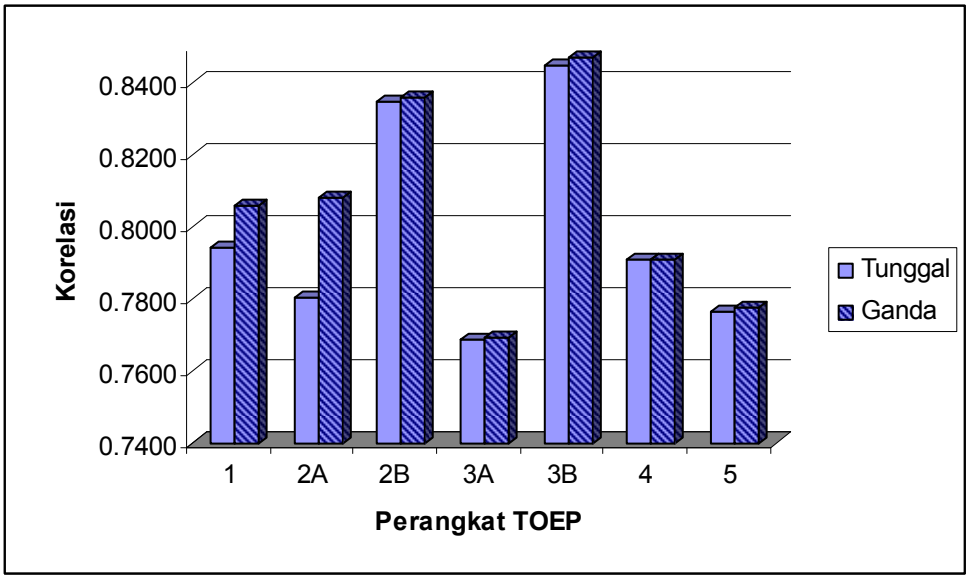
Tabel 2. Hasil Estimasi Koefisien Korelasi dan Kontribusi

Perangkat TOEP	Model			
	$Y=b_0+b_1X$		$Y= b_0+b_1X_1+b_2X_2$	
	r	r^2	r	r^2
1	0.7943	0.6309	0.8060	0.6497
2A	0.7801	0.6085	0.8081	0.6530
2B	0.8349	0.6970	0.8357	0.6984
3A	0.7687	0.5908	0.7692	0.5917
3B	0.8445	0.7132	0.8467	0.7169
4	0.7910	0.6257	0.7910	0.6257
5	0.7765	0.6030	0.7773	0.6043

Keterangan : $\hat{Y}$ skor TOEFL prediksi, X skor TOEP (regresi tunggal)

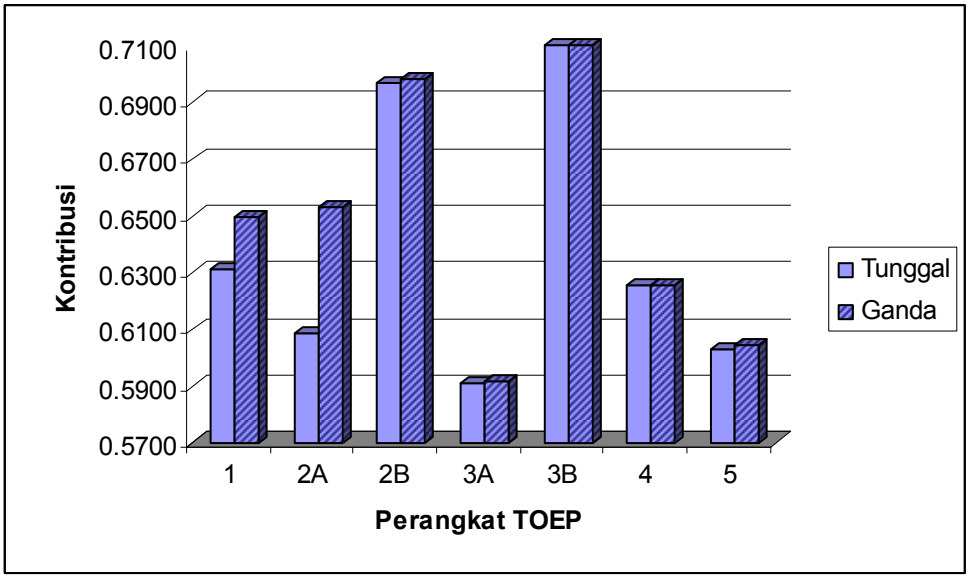
$\hat{Y}$ skor TOEFL prediksi, X_1 skor TOEP *Listening*, X_2 skor TOEP *Reading*

(regresi ganda)



Gambar 3. Korelasi TOEP dan TOEFL dengan Regresi Tunggal dan Ganda

Hasil perhitungan korelasi tersebut menunjukkan kecenderungan bahwa korelasi dengan dua prediktor terhadap TOEFL lebih tinggi dibandingkan korelasi dengan prediktor tunggal. Demikian pula koefisien korelasi determinasi (r^2) yang menunjukkan persentase kontribusi TOEP dalam memprediksi TOEFL. Perbandingan kontribusi pada kedua model disajikan pada Gambar 4.



Gambar 4. Koefisien Determinasi TOEP terhadap TOEFL dengan Regresi Tunggal dan Ganda

Dengan menggunakan data empiris, selanjutnya dapat diestimasi konstanta dan koefisien pada persamaan regresi, yang disajikan pada Tabel 3 untuk model regresi tunggal dan regresi ganda.

Tabel 3. Persamaan Regresi untuk Memrediksi Skor TOEFL dengan Skor TOEP

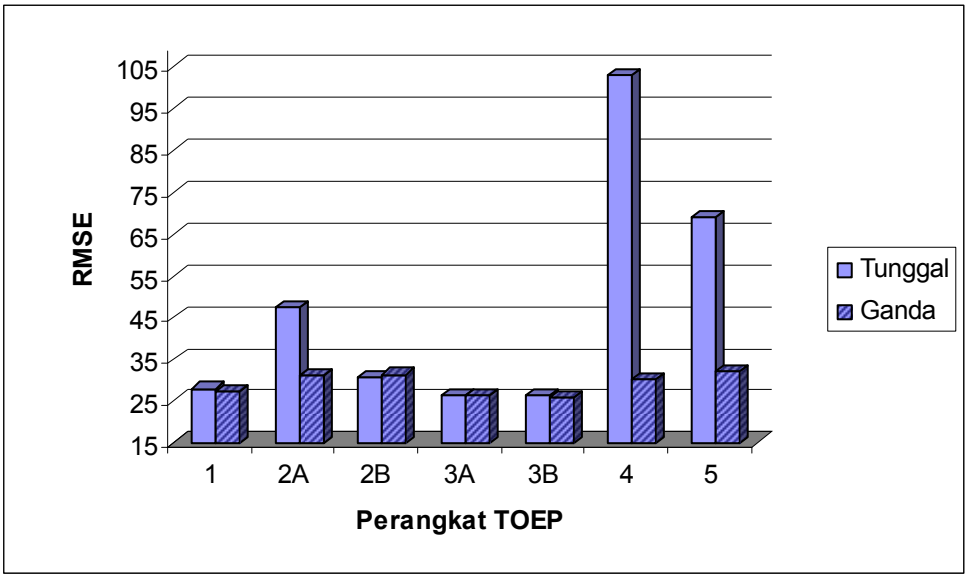
Perangkat TOEP	Persamaan Prediksi (Dengan $\hat{Y}$ skor TOEFL prediksi, X skor TOEP)	Persamaan Prediksi (Dengan $\hat{Y}$ skor TOEFL prediksi, X_1 skor TOEP *Listening*, X_2 skor TOEP *Reading*)
1	$\hat{Y}$ = 3,381 . X + 266,214	$\hat{Y}$ = 274,449+1,285. X_1 +2.401. X_2
2A	$\hat{Y}$ = 4,321 . X + 251,435	$\hat{Y}$ = 264,609+2,977. X_1 +1.120. X_2
2B	$\hat{Y}$ = 4,268 . X + 234,846	$\hat{Y}$ = 239,063+2,273. X_1 +1.922. X_2
3A	$\hat{Y}$ = 3,630 . X + 252,836	$\hat{Y}$ = 254,244+1,917. X_1 +1.692. X_2
3B	$\hat{Y}$ = 3,923 . X + 218,624	$\hat{Y}$ = 223,336+2,210. X_1 +1.634. X_2
4	$\hat{Y}$ = 4,321 . X + 251.435	$\hat{Y}$ = 243,872+1,377. X_1 +1.383. X_2
5	$\hat{Y}$ = 4,268 . X + 234.846	$\hat{Y}$ = 230,464+1,469. X_1 +1.759. X_2

Persamaan regresi pada Tabel 3 tersebut digunakan untuk memprediksi skor TOEP menggunakan skor TOEFL, dengan skor TOEP sebagai prediktor pada model regresi tunggal, dan skor TOEP *Litening* dan skor TOEP *Reading* sebagai prediktor pada model regresi ganda. Dengan menggunakan skor TOEFL sebenarnya dan skor TOEFL hasil prediksi, dihitung MSE dan RMSE dengan menggunakan persamaan 4 dan persamaan 5. Hasilnya disajikan pada Tabel 4.

Tabel 4. MSE dan RMSE pada Regresi Tunggal dan Regresi Ganda

Perangkat TOEP	Model			
	$\hat{Y} = b_0 + b_1 X$		$\hat{Y} = b_0 + b_1 X_1 + b_2 X_2$	
	MSE	RMSE	MSE	RMSE
1	784.6949	28.01241	744.7366	27.28986
2A	2257.273	47.51077	969.1483	31.13115
2B	929.3059	30.48452	983.7762	31.36521
3A	689.0703	26.25015	687.6404	26.2229
3B	684.1966	26.15715	675.3568	25.98763
4	10573.64	102.8282	921.3117	30.35312
5	4755.721	68.96173	1036.608	32.1964

Perbandingan RMSE pada kedua model digambarkan pada Gambar 5.



Gambar 5. Perbandingan RMSE pada Regresi Tunggal dan Regresi Ganda

Mencermati hasil yang disajikan pada Gambar 5 tersebut, dapat diperoleh kecenderungan bahwa RMSE pada model regresi ganda ketujuh perangkat lebih stabil hasilnya dibandingkan dengan RMSE pada model regresi tunggal. Hasilnya juga menunjukkan kecenderungan bahwa RMSE pada model regresi ganda lebih kecil dibandingkan RMSE pada model regresi tunggal. Hal ini menunjukkan bahwa model regresi ganda dengan prediktor skor TOEP *Litening* dan skor TOEP *Reading* lebih akurat untuk memprediksi skor TOEFL dibandingkan dengan menggunakan skor TOEP saja sebagai prediktornya.

Kesimpulan dan Diskusi

Berdasarkan hasil estimasi korelasi, pada korelasi ganda diperoleh hasil lebih tinggi dibandingkan hasil korelasi tunggal. Halini menunjukkan bahwa kontribusi dua variabel prediktor dalam menjelaskan varians kriteria lebih besar dibandingkan dengan hanya menggunakan satu variabel prediktor saja. Semakin tinggi korelasi, variabel prediktor akan semakin akurat memprediksikan variabel kriteria. Keakuratan prediksi dengan mengunakan model ganda ini didukung oleh hasil perbandingan RMSE, pada model regresi ganda dihasilkan RMSE yang lebih kecil dan stabil.

Salah satu penyebab lebih tingginya korelasi pada model ganda yakni adanya muatan multidimensi pada TOEP. TOEP yang dijadikan bahan studi ini terdiri dari 2 komponen kompetensi komunikatif, yaitu Listening dan Reading, yang memiliki konstruk dan sifat yang berbeda. Terkait dengan hal ini, diperlukan penelitian lebih lanjut tentang muatan multidimensi data TOEP.

Penelitian lanjutan tentang *benchmarking* khususnya perbandingan model regresi tunggal dan ganda yang dimanfaatkan pada prediksi skor kriteria dengan skor tertentu perlu dilakukan. Penelitian simulasi dengan memanfaatkan model data tertentu dengan mempertimbangkan variabel panjang tes, banyaknya variabel prediktor, muatan dimensi data, dan pembobotan tiap subtes dapat dilakukan untuk menambah pengetahuan tentang hal-hal yang mempengaruhi hasil estimasi prediksi skor kriteria dengan menggunakan variebel prediktor.

Referensi

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

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.

Cizek, G.J., Rosenberg, S.L. & Koons, H.H. (2008). Source of validity evidence for educational and psychological test. *Educational and Psychological Measurement,* Vol. 68, pp. 397-412.

Direktorat PSMA. 2007. Laporan Pengembangan Test of Englis Proficiency 2007. Dit PSMA Mandikdasemen. Tidak dipublikasikan.

Direktorat PSMA. 2008. Laporan Pengembangan Test of Englis Proficiency 2008. Dit PSMA Mandikdasemen. Tidak dipublikasikan.

Direktorat PSMA. 2009. Laporan Pengembangan Test of Englis Proficiency 2009. Dit PSMA Mandikdasemen. Tidak dipublikasikan.

Fernandes, H. J. X. (1984). *Evaluation of educational program*. Jakarta: National Education Planning, Evaluating and Curriculum Development.

Kerlinger, F.N. (1986). *Asas-asas penelitian behavioral* (Terjemahan L.R. Simatupang). Yogyakarta: Gajahmada University Press.

Kleinbaum, D.G dkk. (1998). *Applied Regression Analysis and Other Multivariate Methods*. Pacific Groove : Duxbury Press.

Lawrence, M.R. (1994). Question to ask when evaluating test. *Eric Digest*. *Artikel*. Diambil dari: http://www. ericfacility. net/ ericdigest/ ed.385607.html tanggal 6 Januari 2007.

Linn, R.L. & Gronlund, N.E. (1995). *Measurement and assessment in teaching* (7th ed.). EnglewoodCliffs, NJ: Prentice-Hall.

Lissitz, W. & Samuelsen, K. (2007). Further clarification regarding validity and education. *Educational Researcher*, Vol. 36, No. 8, pp. 482-484.

Messick, S. (1989). Validity. Dalam R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13-103). New York: Macmillan.

Nunally, J. (1978). *Psychometric theory (2nd ed.)* . New York: McGraw Hill.

Pedhazur, E.J. (1973). *Multiple Regression in Behavioral Research*. New York : Holt, Rinehart and Winston.

Syaifudin Azwar. (2000). *Reliabilitas dan validitas* (Edisi 4). Yogyakarta: Pustaka Pelajar.

Walpole, R.E. dkk. (2002). *Probability and Statistics for Engineers and Scientists*. Upper Saddle River : Prentice-Hall.

Woolfolk, A. E. & McCune, L. N. (1984). *Educational psychology for teachers*. Englewood Cliffs, NJ.: Prentice Hall, In.