JURNAL PSIKOLOGI
1999, No. 2, 93 - 105

GUESSING: INSTRUCTED OR DISCOURAGED PENALIZED OR UNPENALIZED?

Sutarlinah Soekadji
Universitas Indonesia

ABSTRACT

The answers of those questions are found through computerized guessing simulations, using TPA items answered by 1395 examinees in response to instruction that discouraged guessing. The results show that guessing shortened the score effectivity range of the subtests, and increasing average subtest scores, which means increasing the average index of difficulty levels (decreasing item difficulties). Subtest scores penalized for guessing tend to be justly corrected, and subtest scale scores based on norms for guessing option B only, or C only, with or without penalty yield high correlations (.92 to .99) with the original scale scores (without guessing nor penalty). Using random option guessing slightly lowers the correlations (.88 to .97). Guessing penalized or unpenalized yield the same slightly lower alpha reliability (around 0, 03). Thus guessing should be encouraged, especiallyfor the benefit of those who are too anxious to guess, or unwilling to take the risk. Even in the process of scoring, guessing could be provided for blank responses.

Keywords: *Guessing, penalized, unpenalized.*

Menebak *(guessing)* dikenal sebagai sumber kesesatan yang bandel dalam tes kognitif seperti tes prestasi dan tes bakat, sedangkan berlagak *(faking bad/ faking good)* merupakan masalah dalam pengukuran kepribadian/afektif, terutama yang berbentuk inventori. Beberapa cara digunakan untuk mengurangi menebak, misalnya menggunakan rumus koreksi, memberikan petunjuk jelas agar tidak menebak, menggunakan cara penyekoran berbobot khusus. Banyak peserta didik maupun pengajar percaya bahwa menebak menguntungkan penempuh tes, terutama pada tes berbentuk pilihan majemuk *(multiple choice)* sehingga *multiple choice* juga dijuluki *multiple guess* (Mehren & Lehman, 1984). Ada pula yang menganggap *multiple choice* sebagai tes untung-untungan hitung kancing. Bila pendapat-pendapat ini benar (maupun salah), apakah akibatnya bila skor penempuh tes mendapat penalti atau tidak mendapat penalti? Bila menebak dianggap wajar, perlukah menebak ini mendapat peringatan, atau justru disarankan? Masalah inilah yang diteliti dalam studi simulasi komputer menggunakan skor item-item TPA (Tes Potensi Akademik) versi PPS-UII-99-1. Secara rinci penelitian

ISSN : 0215 - 8884

ini berusaha menemukan cara terbaik menghadapi perilaku ”menebak” dalam pelaksanaan TPA, dengan mempertimbangkan akibat-akibat tebakan terhadap karakteristik psikometri skor tes. Simulasi perlakuan menebak bervariasi dari perlakuan "peringatan untuk tidak menebak", diubah dengan menambah jawaban tebakan pada pilihan jawaban B, pada pilihan jawaban C, dan secara random pada jawaban A, B, C, D, dan E, pada item-item yang tidak dijawab.

Mengapa Penempuh Tes Menebak, dan Mengapa harus Kena Penalti?

Pertanyaan sederhana ini perlu diajukan, agar terungkap persepsi mengenai dasar-dasar menebak dan imbalannya. Jawaban sederhananya adalah penempuh tes menebak karena ia tidak memiliki cukup pengetahuan atau kemampuan untuk menjawab dengan tepat. Karena itu ada lebih dari satu kemungkinan cara menebak. Kemungkinan pertama adalah menebak jawaban “asal-asalan” atau membuta, yaitu menjawab secara random di antara pilihan-pilihan yang ditawarkan, dan yang kedua, menebak jawaban berbekal pengetahuan yang dimiliki (penempuh tes menggunakan segala pengetahuan dan kemampuan untuk memilih jawaban yang paling besar kemungkinan benarnya). Kebanyakan mereka yang setuju adanya penalti terhadap tebakan adalah mereka yang mempersepsikan bahwa tebakan dilakukan secara random. Ada kalanya cara ini dilakukan oleh penempuh tes, bila materi yang diteskan jauh di atas kemampuan peserta tes. Pada umumnya, proporsi tebakan asal-asalan ini kecil, sedang dipihak lain percaya bahwa penempuh tes yang termotivasi untuk mengerjakan tes sebaik-baiknya menggunakan segala kemampuannya untuk menjawab dengan benar *(informed guessing).* Salah satu cara mendapat jawaban benar adalah sebanyak mungkin mengeliminasi pilihan-pilihan yang tidak mungkin benar.

Ditinjau dari segi lain, yang tidak setuju menebak menganggap bahwa menebak itu perilaku curang, tidak berbeda dengan intensi untuk menipu, secara moral tidak benar, karena ini semacam perjudian. Lagi pula menebak mempengaruhi aspek-aspek psikometri tes, karena kemasukan varians random keberuntungan dalam menebak. Secara teoretik, dengan ditambahnya varians random ke skor tes akan merendahkan reliabilitas dan validitas (Lord, 1963), meski berbagai penelitian empirik tidak konsisten hasilnya (Blommers & Lindquist, dalam Ebel, 1965; Sabers & Feld, 1968; Troub et al., 1969; Hakstian & Kansup, 1975). Mebren dan Lehmann (1984) menambahkan bahwa dalam kenyataan *informed guessing* tidak merendahkan validitas.

Tes pilihan majemuk kadang-kadang dituduh tidak peka terhadap pengetahuan parsial, karena item diskor 0 atau 1. Jadi, penempuh tes yang memiliki sedikit pengetahuan tetapi tidak cukup untuk memilih jawaban yang benar, mendapat skor 0, sama seperti mereka yang tidak tahu apa-apa. Bila penempuh tes diminta membuat *informed guessing* beberapa item, sebagian tebakannya akan benar, dan ini imbalan bagi pengetahuan parsialnya. Jadi, bila berdasarkan item tunggal tidak ada imbalannya, berdasarkan keseluruhan tes pengetahuan parsial ini ada maknanya.

Mencegah Menebak

Seberapa besarkah keuntungan menebak sehingga harus dicegah? Menebak 30 soal Benar-Salah pada jawaban "Salah" semua, paling besar mendapat angka separuhnya (15). Menebak satu item dengan lima pilihan, kemungkinan benar 1/5, tetapi menebak lima item pada pilihan yang sama (misalnya B semua dari pilihan ABCDE) kemungkinan salah satu item benar mendekati 1. Ini disebabkan biasanya pembuat tes meletakkan jawaban yang benar menyebar merata pada semua pilihan. Meskipun kemungkinan mendapatkan skor sempurna (tebakan benar semua) sangat kecil tetapi tambahan satu skor untuk setiap 5 item hasil menebak ini mungkin dapat berbeda signifikan dengan skor bila tidak menebak. Dalam tes semacam TPA, terutama pada Subtes Kuantitatif dan Penalaran, penyebaran skor begitu besar. Mereka yang kurang mampu dalam kedua subtes itu cenderung tidak sempat melakukan *informed guessing*, sehingga lebih banyak menggunakan *random guessing* (menurut pengalaman penulis sebagai *proctor*, *random guessing* cenderung dilakukan dengan mengisi kolom pilihan tertentu, yaitu pilihan B semua, C semua, atau D semua). Tambahan 1/5 skor tebakan yang benar bagi mereka yang kurang mampu dalam kedua subtes tersebut memperkecil *standard deviasi*, skor-z maupun rangking di antara peserta tes yang tidak menebak, atau yang menebak dengan cara yang berbeda.

Penyelenggara tes yang menganggap menebak sebagai masalah yang harus dicegah, menggunakan petunjuk pada waktu menyajikan tes, agar penempuh tes tidak menebak, dengan penalti pengurangan skor untuk jawaban salah. Meskipun demikian, sejak dulu telah dilakukan penelitian yang menunjukkan bahwa petunjuk penyajian yang menyarankan tidak menebak menimbulkan efek yang berbeda pada penempuh tes, dan perbedaan ini mempengaruhi skor yang didapat. Kepribadian yang dideskripsikan sebagai *"submissive"*. dengan indikator introversi, lebih suka bergumam dari pada mengekspresikan diri, cemas, harga-diri rendah, dan terlalu memperhatikan kesan orang lain terhadap diri mereka (Sherriffs & Boomer, 1954), dan mereka yang tidak berani mengambil resiko (Slakter, 1968), cenderung patuh mengikuti petunjuk "jangan menebak", bahkan lebih baik tidak menjawab meskipun tebakannya bersifat *informed guessing.* Mereka dirugikan dibanding peserta lain yang tidak mempedulikan petunjuk tersebut. Mebren & Lebmann (1984) menyarankan petunjuk semacam itu hendaknya tidak digunakan, kecuali pada (1) tes kecepatan, yang tidak diharapkan semua penempuh tes bisa menyelesaikan semua item, dan (2) tes diagnostik, yang hasilnya digunakan untuk kepentingan penempuh tes sendiri yaitu untuk mengetahui kekuatan dan kelemahan, dengan mengharapkan penempuh tes berusaha keras mencapai skor semaksimal kemampuan yang sesungguhnya.

Rumus Koreksi

Berkaitan dengan petunjuk untuk tidak menebak, penalti atau hukuman dilakukan dengan rumus "koreksi" berikut ini.

$$S = R - W / (A-1)$$

S = Score
R = Right
W = Wrong

A = Alternative

Skor tes sama dengan banyaknya item yang benar, dikurangi banyaknya item yang salah setelah dibagi banyaknya pilihan item dikurangi 1.

Menggunakan rumus tersebut, maka pada item yang memiliki 5 pilihan, bila seseorang menebak 1 kali benar dan 4 kali salah maka nilainya kembali seperti bila ia tidak menebak. Dalam buku persiapan menempuh GRE, Mortison (1990: 23) dalam bahasannya *"The guessing gold mine"* menyarankan penempuh tes untuk menebak, dengan alasan *"you have nothing to lose, and everything to gain*, sedang dalam pelatihan SAT sarannya adalah, *"For totally random guessing, it's no gain, no pain. Blind guessing is not the way to get your best score "* (Mortison, 1988: 23).

Kelemahan adanya penalti tersebut adalah pilihan yang salah tidak selalu merupakan hasil tebakan, dan ini merugikan penempuh tes karena jawaban yang salah merendahkan skor jawaban yang benar. Ada saran lain karena itu bentuknya bukan penaltl tetapi berbentuk "tebusan" bila jawaban dikosongi (tidak ada pilihan jawaban yang ditandai). Rumus tebusan yang dikemukakan Traub et al. (1969) adalah sebagai berikut:

$$S = R + 0/A$$

0 = banyaknya item yang tidak dijawab

Perlakuan ini tampaknya tidak menghambat penempuh tes yang "penakut". Untuk tes kecepatan tebakan menjadi masalah, cara ini sangat efektif menghambat perilaku menebak. perilaku menebak. Dalam hal seperti itu, jelas tidak ada keuntungan untuk menebak secara random item-item yang belum terjawab.

Rumus koreksi muncul dari pengukuran pendidikan. Padahal penggunaan rumus koreksi untuk menebak pada tes pengukuran hasil belajar cenderung *underestimate,* karena pada kebanyakan tes hasil belajar, ada kesempatan untuk menciutkan banyaknya "pilihan jawaban" sebelum ditebak. Bila dari lima pilihan jawaban dapat dieliminasi dua pilihan, maka tersisa tiga pilihan yang kemungkinan mendapat skor benar adalah 1:3, sedang penalti untuk menebak hanya 1/4. Makin banyak pilihan dapat dieliminasi, makin besar peluang untuk mendapatkan skor benar, sebaliknya makin banyak pilihan yang disediakan pada setiap item, makin kecil angka koreksi.

Berlawanan dengan pengaruh menciutkan pilihan adalah adanya distraktor jebakan yang cenderung membuat koreksi untuk menebak *overestimate.* Distraktor jebakan adalah *distractor* (pilihan salah) yang tampak sebagai pilihan yang mungkin sekali paling tepat atau paling benar. Pilihan semacam itu seringkali menarik sehingga siswa yang benar-benar tahu pun terjebak. Menurut Price (1964, dalam Nunnally,1978), koreksi untuk tebakan meskipun demikian pada umumnya cenderung *undercorrect* bukan *overcorrects.*

Menurut Nunnally (1978), pada beberapa tipe pengukuran pendidikan, ada dua alasan untuk tidak memberi petunjuk agar subjek mengerjakan semua item. Pertama, petunjuk ini cenderung menurunkan reliabilitas skor tes. Penurunan reliabilitas cenderung kecil, tidak lebih dari 0,03 atau 0,04. Perbedaan reliabilitas sebesar itu mungkin tidak penting untuk pengukuran dalam penelitian dasar, tetapi mungkin penting untuk pengukuran dalam

pendidikan. Untuk tujuan praktis juga sering lebih baik menambah reliabilitas dengan menambah banyaknya item daripada mencegah siswa menebak untuk menjawab semua item. Alasan kedua adalah petunjuk tersebut mungkin menjerumuskan siswa ke arah sikap yang jelek, yaitu “mengawur" dalam mengerjakan tugas sehari-hari. Padahal, seharusnya siswa dididik untuk menyelidiki kenyataan dan memikirkan jawaban untuk masalah yang harus diselesaikan, sehingga memaksa menebak pada tes pilihan majemuk akan menimbulkan kebiasaan intelektual yang jelek. Nunnally meragukan kebenaran pendapat tersebut, apakah benar menebak dalam tes pilihan majemuk akan sejauh itu merasuk dalam pikiran siswa. Siswa mungkin bereaksi negatif terhadap petunjuk untuk menebak selama tes, tetapi apakah benar kebiasaan belajar dan berpikir benar-benar terpengaruh.

Sebelum ditemukan rumus koreksi yang memuaskan, menurut Nunnally (1978) sebaiknya digunakan rumus koreksi untuk tebakan berdasar tebakan membuta. Skor-skor yang terkoreksi biasanya memiliki reliabilitas sama dengan skor-skor yang tidak dikoreksi pada tes yang suka, sama meski berdasar penelitian (Guilford, 1954; Lord, 1963; Price, 1964, dalam Nunnally, 1978) validitas prediktif skor yang dikoreksi lebih tinggi sekitar 0,03. Sebaiknya digunakan petunjuk agar semua item dijawab, kecuali bila ada alasan lain**,** yang mendukung larangan menebak. Petunjuk seperti itu dapat dilakukan pada hampir semua pengukuran yang dipakai dalam penelitian dasar, dan juga dalam kebanyakan tes pendidikan. Bahkan bila perlu dapat dijelaskan bahwa mereka mendapat bonus 1 skor bagi k item yang dikosongkan (k adalah banyaknya pilihan pada satu item).

Pengaruh Tebakan terhadap Hasil Analisis Item

Istilah analisis item ditujukan untuk menyebut sekelompok statistik yang dapat dihitung untuk setiap item dalam tes. Analisis item yang baik sering sangat informatif bila tes tidak reliabel atau bila tes gagal mencapai tingkat validitas tertentu. Analisis item dapat menunjukkan *mengapa* tes reliabel (atau tidak reliabel), dan dapat membantu memahami mengapa skor tes dapat digunakan untuk memprediksikan kriteria tertentu, tetapi tidak untuk kriteria yang lain. Item analisis juga dapat menyarankan cara untuk meningkatkan karakteristik pengukuran suatu tes. Hal ini disebabkan, tes kadang-kadang reliabilitas dan validitasnya terbatas karena berisi item-item yang susunan kalimatnya kurang baik atau pertanyaannya merupakan jebakan yang membutuhkan keahlian berpikir yang berbelit-belit. Sebaliknya, item dapat kelihatan bertampang bagus *(face validity-nya* tinggi) tetapi sebenarnya tidak mengukur konstruk ranah yang dirancang untuk diukur. Validitas maupun reliabilitas tes dapat diperbaiki dengan membuang item-item semacam itu. Langkah semacam ini tampak kontradiktif, sebab proposisi dasar teori reliabilitas adalah makin banyak itemnya makin tinggi reliabilitasnya (dengan syarat item-item itu mengukur hal yang sama). Jadi sebenarnya ini tidak berlawanan, karena yang digugurkan adalah item yang jelek, dalam arti tidak mengukur hal yang sama (Murphy & Davidshofer, 1994).

Tiga hal pokok yang dilakukan dalam analisis item, yaitu menjawab tiga pertanyaan. Pertanyaan yang wajar dalam menghadapi tes pilihan majemuk adalah "Berapa banyak orang yang memilih tiap-tiap pilihan?". Pertanyaan ini dijawab menggunakan "analisis distraksi" (distraksi adalah pilihan yang mengalihkan perhatian dari pilihan yang benar). Pertanyaan wajar kedua adalah "Berapa banyak yang memilih jawaban benar dibanding yang memilih jawaban salah?". Pertanyaan ini dijawab dengan analisis yang menghitung "taraf kesulitan". Pertanyaan yang terakhir adalah "Apakah jawaban terhadap item ini ada hubungannya dengan jawaban terhadap item-item lain?" Jawaban dari pertanyaan ini berkaitan dengan analisis "daya diskriminasi item"(Murphy & Davidshofer, 1994).

Masuknya varians *guessing*, maka berubahlah frekuensi pilihan masing-masing distraktor, terutama pada item-item yang sulit. Demikian pula taraf kesulitan meningkat (dalam arti lebih mudah) bila kebetulan *guessing* mengena pada pilihan jawaban yang benar dan tidak berubah bila tidak mengena. Akibat yang lain adalah korelasi item dengan total berubah karena skor total bertambah dengan skor hasil tebakan, demikian pula bila tebakan "mengena" skor item berubah dari 0 menjadi 1. Dengan melakukan penalti bagi item yang salah, maka skor item yang salah tebak menjadi minus. Taraf kesulitan item sulit yang banyak ditebak dengan tebakan yang tidak mengena menjadi makin kecil (item menjadi makin sulit). Korelasi item total didapat dari skor item yang lebih tinggi variasinya (rentangnya dari -1/[k-1] sampai 1, bukan antara 0 sampai 1), dikorelasikan dengan skor tes yang telah dikoreksi diperoleh korelasi makin tinggi untuk skor yang sulit dan makin rendah untuk skor yang mudah. Ini berarti "daya diskriminasi" item makin tampak bila dilakukan penalti/koreksi terhadap tebakan.

Homogenitas dan Reliabilitas Alpha

Analisis konsistensi internal merupakan usaha menentukan derajad interrelasi antara item-item. Dalam istilah operasional hal ini mempertanyakan apakah skor setiap item saling berkorelasi. Bila semua skor item dalam satu tes saling berkorelasi positif, maka tes tersebut homogen. Jadi, homogenitas suatu tes dapat didefinisikan sebagai konsistensi kinerja keseluruhan item dalam tes (Murphy & Davidshofer, 1994). Homogenitas secara tidak langsung sama dengan daya beda item. Item yang tidak dapat membedakan mereka yang skor totalnya tinggi dengan yang rendah, adalah item-item yang tidak homogen dengan item-item lain.

Reliabilitas dinyatakan dalam rata-rata korelasi antar item. Tampak di sini bahwa tes makin homogen bila rata-rata korelasi antar item tinggi yaitu bila item cenderung mengukur trait yang sama. Koefisien alpha Cronbach adalah salah satu ukuran homogenitas yang terkenal yang dapat diperoleh dari rumus berikut ini:

$$rkk = \frac{k}{k-1}\left[1 - \frac{\sum St^2}{Sx^2}\right]$$

k adalah banyaknya item dalam tes, ΣSt^2 adalah jumlah rata-rata varians skor-skor item, dan Sx^2 adalah varians skor item (Brown, 1976). Koefisien alpha dapat diinterpretasikan sebagai rata-rata korelasi antara suatu tes dengan tes lain yang sama

panjangnya yang diambil dari ranah yang sama.

Sasaran teori reliabilitas adalah untuk mengukur estimasi *error* dalam pengukuran dan menemukan saran untuk meningkatkan mutu tes sehingga *error* tersebut sekecil mungkin (Murphy & Davidshofer, 1994). Bila korelasi item-total berubah akibat *guessing*, maka bertambah besar varians *error*. Dengan demikian, reliabilitas alphanya akan lebih rendah. Sebaliknya, bila dilakukan koreksi ada kemungkinan *error* akibat *guessing* digantikan oleh *error* akibat koreksi. Akibat menebak, estimasi rata-rata skor tes akan meningkat. Banyaknya peningkatan berbanding terbalik dengan banyaknya item yang benar seandainya penempuh tidak menebak, sebab makin sedikit pengetahuan yang dimiliki semakin banyak orang menebak. Sebaliknya, bila tebakan merupakan faktor yang diperhitungkan, perlu dipikirkan mengenai rentang skor efektif, bukan skor yang mungkin dicapai. Rentang skor efektif dengan mudah dapat didefinisikan dalam kasus hipotetik dimana seorang subjek menjawab semua item dengan menebak dan bila subjek menjawab semua item benar. Dalam hal seperti ini batas bawah rentang efektivitas skor adalah banyaknya item tes dikalikan dengan kemungkinan benar dalam menebak. Misalnya, bila 40 item dengan empat pilihan jawaban, rentang efektivitas skor adalah 10 sampai 40. Bila model untuk menebak membuta benar, maka siswa yang mendapat skor 10 adalah siswa yang tidak tahu apapun mengenai materi yang diteskan. Skor di bawah 10 adalah skor yang terjadi hanya karena kebetulan *(by chance).* Ini dapat terjadi misalnya, bila tes berisi item-item yang banyak perang-kapnya yang menyebabkan individu dengan kemampuan yang sangat rendah mencapai skor yang lebih jelek lagi.

Tujuan Penelitian

Penelitian ini berusaha menjawab persoalan-persoalan apakah ada perbedaan (1) rentang efektivitas skor subtes, (2) taraf kesulitan subtes-subtes, (3) homogenitas item-item, (4) reliabilitas alpha subtes-subtes, (5) skor skala subtes yang diperoleh, dalam kaitannya dengan perlakuan terhadap jawaban item-item TPA versi PPS-UI-99-1, dalam bentuk simulasi. Perlakuan merupakan kombinasi antara perlakuan tanpa tebakan dan tiga macam tebakan (tebakan pertama: item-item yang tidak dijawab diberi jawaban pilihan B semua; kedua: diberi jawaban pilihan C semua; dan ketiga: diberi jawaban pilihan secara random A sampai E), dikombinasikan dengan penyekoran memakai rumus koreksi dan tanpa rumus koreksi.

METODE PENELITIAN

Populasi penelitian ini adalah peserta ujian masuk berbagai program yang dikelola oleh Program Pascasarjana Universitas Indonesia. Sampel yang dipakai berasal dari 1395 orang, yaitu mereka yang menempuh ujian masuk Gelombang I untuk tahun ajaran 1999/2000.

Penelitian ini menggunakan simulasi komputer terhadap data jawaban item-item TPA versi PPS-UI-99-1. TPA tersebut terdiri dari tiga subtes: Verbal, Kuantitatif, dan Penalaran. Masing-masing subtes terdiri dari 50 item dengan lima pilihan

jawaban (A, B, C, D, dan E). Alokasi waktu 50 menit per subtes. Item-item yang dipakai pada setiap subtes dua macam: (1) sebelum digugurkan (semua item diperhitungkan), khususnya untuk mengukur rentang efektivitas skor, rata-rata kesulitan item, dan homogenitas item (korelasi item-total); (2) digunakan item-item yang homogenitasnya memenuhi syarat pada berbagai perlakuan.

Korelasi item-total dan reliabilitas dihitung menggunakan perintah RELIABILITY *(SPSS/ Versi 7.5 for Windows).* Item-item digugurkan bila pada angka yang tercantum dalam kolom *"Alpha if item deleted'* lebih besar daripada alpha subtes yang diperoleh. Rentang efektivitas skor dihitung dengan cara mendapatkan skor terendah yang mungkin dicapai pada semua perlakuan. Skor tertinggi yang mungkin dicapai sama dengan banyaknya item. Rentang efektivitas skor ini dibandingkan dengan rentang yang diperoleh secara hipotetik menggunakan sederet data yang hanya dijawab B, dijawab C, dan dijawab secara berurutan ABCDE, dan diskor berdasarkan kunci baik tanpa rumus koreksi maupun dengan rumus koreksi. Taraf kesulitan subtes diperoleh dengan membagi rata-rata skor subtes dengan banyaknya item.

Perbedaan reliabilitas alpha dihitung angkanya tanpa melihat z_r, dari Fisher (Ferguson, 1976:184). Skor skala subtes dihitung menggunakan *mean* dan SD yang didapat berdasarkan perlakuan, yaitu kombinasi antara perlikuan tanpa tebakan, dan tiga macam tebakan, dengan memakai koreksi dan tanpa koreksi. Perbedaan skor skala subtes dihitung menggunakan perintah T-TEST berpasangan *SPSS versi 7.5 for Windows* setelah diuji normalitas dan homogenitas varians. Dari perintah ini juga didapat korelasi berpasangan antara skor skala yang dibandingkan.

HASIL PENELITIAN

Meskipun pada petunjuk tidak disarankan untuk menebak, beberapa peserta tetap melakukan tebakan membuta, yang ditandai dengan mengisi semua jawaban pada bagian akhir dengan pilihan jawaban yang sama (B semua, C semua, atau D semua). Dari banyaknya item yang dijawab bisa diberikan instruksi larangan menebak, diperoleh deskripsi data yang disajikan pada Tabel 1. Dalam perhitungan itu lima subjek tidak diikutsertakan karena ada subtes yang sama sekali tidak dikerjakan. Subtes 1 (Kode TPA-1) rata-rata dijawab 44,41 item, minimum dijawab 7 item, maksimum semua item dijawab. Subtes ini, berupa subtes verbal adalah subtes yang paling mudah di antara ketiga subtes. Pada tabel terlibat bahwa Subtes 2 (TPA-2) dan Subtes 3 (TPA-3) jauh lebih sulit, karena rata-rata lebih dari 20 item yang dikosongi (tidak dijawab), dengan rentang penyebaran yang lebih luas. Terbatasnya kemampuan dan alokasi waktu tes, terutama pada Subtes 2 dan 3, maka dapat diperkirakan bahwa bila disarankan untuk menebak, sebagian tebakan adalah tebakan membuta.

Tabel 1. Deskripsi Item-item yang Dijawab pada Setiap Subtes (k* = 50)

Subtes	N	Minimum	Maximum	Mean	Std. Deviation
TPA-1	1390	7.00	50.00	44.4165	6.3347
TPA-2	1390	2.00	50.00	28.1453	10.6270
TPA-3	1390	1.00	50.00	30.0086	10.9895

*k=banyaknya item per subtes, juga berlaku untuk tabel-tabel berikutnya.

Tabel 2. Deskripsi Item-item yang Benar pada Setiap Subtes dan Taraf Kesulitan Subtes (k = 50)

Subtes	N	Minimum	Maximum	Mean	Std. Deviation	Taraf Kesulitan
TPA-1	1390	2.00	45.00	26.7719	7.1028	.5354
TPA-2	1390	.00	43.00	13.9978	6.9926	.2798
TPA-3	1390	.00	41.00	13.7345	6.1606	.2746

Taraf kesulitan Subtes 2 dan 3 lebih tampak lagi bila item-item telah diskor (Tabel 2). Dari 28 item Subtes 2 yang dijawab, yang benar hanya separuhnya, bahkan dari Subtes 3, yang 8,3 (Tabel 4). Perlakuan dengan koreksi mengembalikan ke skor tanpa tebakan, bila dirata-ratakan cenderung sedikit *overcorrected* tetapi tidak cukup signifikan (Tabel 3 dan 4 dengan koreksi). Akibat rumus koreksi skor Subtes 1 rata-rata menurun menjadi 1,83 untuk 40 item, sedangkan sebelum tebakan skor terendah 2, demikian juga skor subtes 2 dan 3, yang rata-rata lebih rendah dari nol.

Taraf kesulitan yang diperoleh dari data setelah item-item yang tidak homogen digugurkan, sebelum ditambah tebakan dan dikoreksi berturut-turut dan Subtes 1 sampai 3 adalah 0,53; 0,28; dan 0,27 (Tabel 5C). Setelah ditambah jawaban tebakan B tanpa koreksi skor subtes-subtes meningkat, sehingga taraf kesulitan naik menjadi 0,56; 0,37; dan 0,37; dengan koreksi kenaikan taraf kesulitan menyusut menjadi 0,48; 0,26, dan 0,23 (Tabel 5D). Kenaikan pada Subtes 2 dan 3 tanpa koreksi cukup tinggi karena peluang menebak lebih banyak daripada pada Subtes 1. Sama halnya penyusutan setelah dikenai rumus koreksi, peluang koreksi juga lebih tinggi.

Homogenitas item-item juga mengalami perubahan. Item-item yang homogen cenderung lebih tinggi korelasi item-totalnya bila ditambah tebakan. Sebaliknya item-item yang semula rendah korelasi item-totalnya, cenderung menjadi lebih rendah lagi bila diberi tebakan. Sedang item-item yang terletak di antara kedua kondisi homogenitas itu berubah menjadi lebih kecil dan lebih besar tergantung pada kecenderungan homogenitasnya. Reliabilitas alpha mengalami penurunan bila diberi tebakan. Pada perhitungan permulaan, sebelum item-item yang tidak homogen digugurkan, penurunan untuk Tebakan B, tanpa koreksi dari 0,8288 ke 0,7945 atau 0,0343 untuk Subtes 1, tetapi untuk Subtes 2 dari 0,8533 menjadi 0,7994

atau 0,0539, dan untuk Subtes 3 sebesar 0,0636. Reliabilitas alpha tidak berubah dari reliabilitas skor tanpa koreksi untuk perlakuan yang mendapat tebakan, benar kurang dari separuhnya. Taraf kesulitan Subtes 1 adalah 0,53, taraf kesulitm Subtes 2 adalah 0,28, dan taraf kesulitan Subtes 3 sama dengan 0,27. Bila semua item ditebak dan tidak dikenai rumus koreksi, maka didapat skor hipotetik terendah yang disajikan pada Tabel 3.

Tabel 3. Deskripsi Skor Hipotetik terendah Setiap Subtes bils Semus Item Hasil Tebakan (k = 50)

Tebakan	Skor Seluruh-Item-Tanpa-Koreksi			Skor Seluruh-Item-Pakai-Koreksi		
	B	C	ABCDE	B	C	ABCDE
TPA-1	12.00	10.00	11.00	2.50	1.25	1.25
TPA-2	11.00	10.00	12.00	.00	.00	-3.75
TPA-3	11.00	7.00	11.00	1.25	2.50	1.25

Tabel 4. Deskripsi Skor Hipotetik Terendah Setiap Subtes (k = 40)

Tebakan	Skor Seluruh-Item-Tanpa-Koreksi			Skor Seluruh-Item-Pakai-Koreksi		
	B	C	ABCDE	B	C	ABCDE
TPA-1	9.00	6.00	10.00	1.25	2.50	1.25
TPA-2	10.00	8.00	7.00	-2.50	0.00	-2.50
TPA-3	9.00	6.00	8.00	2.50	-1.25	.00

Bila petunjuk menyarankan tidak menebak, skor terendah untuk Subtes 1, 2, dan 3 hanya mendapat 2, 0, dan 0. Tetapi bila diberi tebakan dengan melibatkan seluruh item, skor terendah berubah menjadi 12, 11, dan 11 untuk tebakan B, untuk tebakan C paling rendah: 10, 10, 7, atau rata-rata 9. Terlihat bahwa menggunakan tebakan tanpa koreksi mempersempit rentang efektivitas skor subtes. Rentang efektivitas skor kembali melebar bila digunakan koreksi. Pada perlakuan hanya memakai item-item yang homogen di setiap subtes, rentang efektivitas skor Subtes 1 semula dari 0 sampai 40, dengan berbagai tebakan tanpa memakai koreksi secara hipotetik bila dirata-ratakan ketiga skor minimalnya adalah:

Tabel 5. Mean, SD, dan Taraf Kesulitan Subtes (N=1395)

A. Tanpa Koreksi (k = 50)												
Subtes	TANPATEBAKAN			TEBAKAN-B			TEBAKAN-C			TEBAKAN-ABCDE		
	Mean	SD	TK	Mean	SD	TK	Mean	SD	TK	Mean	SD	TK
TPA-1	26.69	7.23	.53	27.77	6.66	.56	27.90	6.73	.56	28.23	6.81	.56
TPA-2	13.98	6.99	.28	18.57	6.22	.37	17.57	6.42	.35	18.72	6.59	.37
TPA-3	13.73	6.17	.27	18.46	5.59	.37	15.88	5.65	.32	18.29	5.89	.37

B. Pakai Koreksi (k = 50, N= 1 395)												
Subtes	TANPATEBAKAN			TEBAKAN-B			TEBAKAN-C			TEBAKAN-ABCDE		
	Mean	SD	TK	Mean	SD	TK	Mean	SD	TK	Mean	SD	TK
TPA-1	22.30	8.37	.44	22.21	8.33	.44	22.37	8.41	.45	22.78	8.51	.55
TPA-2	10.45	7.85	.21	10.70	7.77	.21	9.46	8.02	.19	10.91	8.23	.21
TPA-3	9.66	6.93	.19	10.57	6.99	.21	7.34	7.06	.14	10.25	7.36	.20

C. Tanpa Koreksi (k =40, N=1395)												
Subtes	TANPATEBAKAN			TEBAKAN-B			TEBAKAN-C			TEBAKAN-ABCDE		
	Mean	SD	TK	Mean	SD	TK	Mean	SD	TK	Mean	SD	TK
TPA-1	22.40	6.57	.56	23.37	6.07	.58	22.98	6.21	.57	23.69	6.21	.59
TPA-2	11.82	6.47	.30	16.20	5.81	.41	14.63	5.95	.41	14.88	6.05	.37
TPA-3	11.99	5.79	.30	15.44	5.34	.39	13.84	5.42	.35	15.67	5.72	.39

D. Pakai Koreksi (k =40, N=l 395)												
Subtes	TANPATEBAKAN			TEBAKAN-B			TEBAKAN-C			TEBAKAN-ABCDE		
	Mean	SD	TK	Mean	SD	TK	Mean	SD	TK	Mean	SD	TK
TPA-1	19.14	7.63	.48	19.21	7.59	.48	18.72	7.76	.46	19.61	7.77	.49
TPA-2	8.84	7.29	.22	10.25	7.26	.26	8.28	7.44	.21	8.60	7.57	.22
TPA-3	8.78	6.62	.22	9.30	6.67	.23	7.30	6.77	.18	9.59	7.15	.24

sedang yang tidak diberi tebakan, reliabilitas menurun sekitar 0,002 setelah dikoreksi. Setelah item-item yang tidak homogen digugurkan, reliabilitas meningkat, dan penurunan akibat tebakan sekitar sama besarnya dengan penurunan pada perhitungan sebelum item-item digugurkan.

Tabel 6. Hasil Perhitungan Reliabilitas Alpha Subtes-Subtes (N=1395)

ALPHA k = 50, TANPA KOREKSI				
	TANPA TEBAKAN	TEBAKAN-B	TEBAKAN-C	TEBAKAN-ABCDE
TPA -1	.8288	.7945	.8005	.7966
TPA -2	.8533	.7994	.8162	.7789
TPA -3	.7965	.7339	.7492	.7172

ALPHA k=50, PAKAI KOREKSI				
	TANPA TEBAKAN	TEBAKAN-B	TEBAKAN-C	TEBAKAN-ABCDE
TPA -1	.8128	.7945	.8005	.7966
TPA -2	.8469	.7994	.8162	.7789
TPA -3	.7813	.7339	.7492	.7172

ALPHA k =40, TANPA KOREKSI				
	TANPA TEBAKAN	TEBAKAN-B	TEBAKAN-C	TEBAKAN-ABCDE
TPA -1	.8452	.8150	.8244	.8152
TPA -2	.8577	.8082	.8222	.7955
TPA -3	.8095	.7647	.7784	.7592

ALPHA k =40, PAKAI KOREKSI				
	TANPA TEBAKAN	TEBAKAN-B	TEBAKAN-C	TEBAKAN-ABCDE
TPA -1	.8436	.8150	.8244	.8152
TPA -2	.8567	.8082	.8222	.7955
TPA -3	.8076	.7647	.7784	.7592

Perhitungan korelasi *Product Moment* antara skor-skala subtes-subtes tanpa tebakan tanpa koreksi dengan skor skala subtes-subtes yang diperoleh dengan cara-cara lain menunjukkan bahwa pada Subtes 1 diperoleh korelasi tinggi antara 0,92 dengan 0,98, pada Subtes antara 0,97 dengan 0,99, pada Subtes 3 antara 0,88 dengan 0,98. Korelasi terendah terdapat pada Subtes 3 dengan tebakan pilihan A, B, C, D, dan E). Sedang perhitungan perbedaan mean menggunakan t-test berpasangan menunjukkan tidak ada perbedaan yang signifikan antara skor skala tanpa-tebakan-tanpa-koreksi dengan skor skala lainnya pada masing-masing Subtes.

Diskusi

Hasil penelitian menunjukkan bahwa tidak semua subjek mematuhi petunjuk untuk tidak menebak. Padahal simulasi menunjukkan bahwa menebak lebih menguntungkan dari pada tidak menebak. Dengan demikian, petunjuk untuk tidak menebak merugikan mereka yang patuh. Akibat tebakan berikutnya adalah skor subtes-subtes meningkat, sehingga taraf kesulitan subtes juga meningkat (tes lebih mudah) dan reliabilitas alphanya lebih rendah antara 0,03 dengan 0,05 dibanding sebelum diberi tebakan. Memakai koreksi maupun tanpa koreksi reliabilitas alpha subtes yang diberi tebakan tidak berbeda. Dengan kata lain, koreksi tidak menurunkan atau menaikkan reliabilitas alpha.

Ditambahkan tebakan rentang efektivitas skor menjadi lebih sempit, tetapi rumus koreksi mengembalikan ke rentang semula. Demikian pula skor Skala memakai-tebakan-memakai-koreksi memiliki korelasi tinggi dengan skor tanpa-tebakan-tanpa-koreksi. Selain itu tidak ada perbedaan yang signifikan antara skor Skala tanpa-tebakan-tanpa-koreksi dengan memakai-tebakan-memakai-koreksi bila skor Skala diperoleh menggunakan norma berdasarkan perhitungan perolehan skor masing-masing perlakuan. Ini berarti tidak ada yang dirugikan bila semua penempuh tes diberi kesempatan untuk menebak.

Saran-Saran

Berdasar hasil penelitian tersebut di atas dalain penyajian tes yang mirip TPA disimpulkan *guessing should be encourage*

and it doesnot matter penalized or unpenlized.

Disarankan agar:

1. Petunjuk penyajian tidak perlu melarang menebak, bahkan disarankan untuk mencoba menjawab semua soal,
2. Perhitungan skor hendaknya menambahkan jawaban tebakan bila penempuh tes mengosongi jawaban.
3. Rumus koreksi dapat digunakan, meski hasil perhitungan memakai atau tanpa memakai rumus koreksi skor skalanya tidak berbeda..

DAFTAR PUSTAKA

Brown, F.G. 1976. *Principles of Educational and Psychological Testing.* Edisi ke-2. New York: Holt, Rinehart and Winston.

Ebel, R.L. 1965. *Measurement of Educational Achievement.* Englewood Cliff, NJ: PrenticeHall.

Ferguson, G. A. (1976). *Statistical Analysis in Psychology & Education.* New York: McGraw-Hill.

Hakstian, A.R. & Kansup, W. (I 975). A comparison of several methods of assessing partial knowledge in multiple-choice tests: II. Testing procedure. *Journal of Educational Measurement, 12,* 231-240.

Lord, F.M. 1963. Formula scoring and validity. *Journal of Educational Measurement, 14,* 33-38.

Mehren, W.A. & Lehmann, I.J. 1984. *Measurement and Evaluation in Education and Psychology.* Edisi ke-3. New York: Holt, Rinehart and Winston.

Mortison, T. H. 1980. *Super Course for the GRE.* New York: Arco.

Mortison, T. H. 1988. *Super Course for the SAT* New York: Arco.

Murphy, K.R. & Davidshofer, C.O. 1994. *Psychological Testing: Principles and* A*pplication.* Edisi ke-3. London: Prentice-Hall International.

Nunnally, J.C. 1978. *Psychometric Theory.* New Delhi: Tata McGraw-Hill.

Sabers, D.L. & Feld, L.S. 1968. An empirical study of the effect of correction for chance success on the reliability and validity of an aptitude test. *Journal of Educational Measurement, 5,* 251-256.

Sherriffs, A.C.& Boomer, D.S. 1954. Who is penalized by the penalty for guessing?*Journal of Educational Psychology, 45,* 81-90.

Slakter, M.J. 1968. The penalty of not guessing. *Journal of Educational Measurement, 5,* 217-221.

Troub, R.E., Hambleton, R.K., and Singh, B. 1969. Effect of promised reward and threatened penalty on performance on a multiple-choice vocabulary test. *Educational dan Psychological Measurement, 29,* 847-861.