BAB V
METODE TES BAHASA

Metode-metode digunakan untuk membangun test tapi bukan pada test itu sendiri. Meskipun mungkin untuk membicarakannaik tidaknya satu test tapi nyatanya tidak bagi metode. Prosedur pilihan ganda mungkin menghasilkan test yang valid vada suatu realisasi tapi tidak padarealisasi lain. Ini merupakan kasusu untuk semua metode dan akan mengingatkan kembali pembahasan selanjutnya terhadap keuntungan dan keruguan potensial dari metode-metode yang berbeda berikut ini.

Pendekatan-pendekatan yang berbeda terhadap ujian bahasa adalah diuraikan di dalam bab I, referensi dibuat bagi efek yang memungkinkan dari metode test terhadap skor test. Terdapat bukti yang sama dalam literatur (lih. Murphy, 1978, 1988; Porter, 1983; Weir, 19…; Boniakowskan, 1986; Alderson dan Urguhast, 1985a) bahwa format test mungkin mempengaruhi ferformansi siswa, membatasi tingkat pengetahuan mengenai efek format test terhadap satu-satunya pendekatan praktis saat ini yaitu untuk melindungi kemungkinan efek format yang mengmbengkan dasar test dengan lebih luas melalui penggunaan variasi format yang valid, praktis, dan reliable bagi ujian tiap keterampilan.kita belum membahas beberapa metode test yang berbeda secara detail pada penggunaanya. Oleh karena itu bab ini memberikan laporan singkat tentang jenis-jenis utama format test dan beberapaa *Manfaat* dan *Kerugian* potensialnya. Ini dimaksudkan untuk memberikan referensi pada pedoman untuk kontruksi test yang akan datang.

Sebagaimana kaidah umum, ia adalah yang terbaaik untuk menilai dengan menggunakan fariatif format tersebut, skor diambil adalah skor gabungan untuk melaporkan tujuan. Syarat atau ketentuan utama untuk ujian menggunakan kerangka komunikaatif adalah bahwa tugas test akan menunjukkan pengolahan percakapan nyata sejauh mungkin dan mencakup susunan keterampilan-keteramilan yang mungkin yang telah diperkenalkan (lih. Appendix I sebagai contoh dari pendekatan ini dalam test TEEP). Penting bahwa test-test yang

dikembangkan dalam paradigma ini akan mempunyai efek washback yang kuat terhadap praktek di kelas bahasa.

5.1 Pendahuluan

5.1.1 Pertanyaan-pertanyaan Pilihan Ganda (MCQS)

Saran untuk penyusun soal-soal pilihan ganda juga dapat dipakai pada penyusunan test-test pemahaman menyimak, struktur, dan osa kata. Ini semua akan dibahas dalam bab ini.

Soal ujian pilihan ganda biasanya diberikan dengan cara di mana kandidat harus memilih jawaban dari pilihan-pilihan yang diberikan, hanya satu yang bear. Proses penilaian semuanya objektif karena penilai tidak diperbolehkan untuk membuat pertimbangan ketika menilai jawaban kandidat, mufakat telahdicapai seperti terhadap jawaban yang benar dari tiap soal. Pemilihan dan penyusunan soal merupakan proses yang subjektif dan keputusan tentang mana jawaban yang benar merupakan persoalan pertimbangan yang subjektif pada penulis.

Manfaat

1. Dalam test pilihan ganda terdapat reliabilitas penialia yang hampir sempurna. Nilai-nilai yang ada pada format subjektif, tak bisa dipengaruhi oleh pertimbangan personal atau oleh keistimewaan-keistimewaan penilai. Selain dapat dipercaya, penilaiannya itu sederhana, lebih cepat dan sering lebih efektif daripada bentuk lain dari test tertulis.
2. Karena soal itu bisa diuji coba dengan hampir mudah, biasanya memungkinkan untuk memperkirakan dulu tingkat kesulitan dari tiap soal. Free-test juga memberikan informasi tentang tingkat yang tiap soal tambah secara fositif terhadap apa yang test pada umumnya nilai. Dwiarti dalam susunan kata-kata dari soal-soal mngkin juga diungkapkan dalam analisis data free-test dan bisa dijelaskan atau dijernihkan dalam tets yang sebenarnya.
3. Format dari soal test pilihan ganda yaiitu seperti tujuan-tujuan dari penyusun test itu jelas dan tegas; para kandidat mengetahui apa yang mereka perlakukan. Pada format-format open-ended dwi arti dalam susunan kata-kata

pada pertanyaan mungkin kadang-kadang menuntun para kandidat yang mengatur waktu tambahan menjwab pertanyaan berbeda dengan apa yang diharapkan penguji.

4. Pada format open-ended, contohnya pertanyaan-pertanyaa jawaban singkat, calon atau pesertaharus menyebutkan keterampilan menulis. Tingkat yang mempengaruhi pengukuran-pengukuran yang akurat terhadap ciri bawaan dinilai bukan ditentukan. Test-test pilihan ganda menghindari kesulitan khusus ini.

Kerugian

1. Bagamanapun juga terdapa bebrapa masalah yang brhubungan dengan penggunaan format ini. Jika seorang kandidat menjawab soal pilihan ganda karena bebrapa kekurangan dalam pertanyan, lembar jawaban tempat ia menjawab jawabannya tidak akan mengungkapkan kebenaran ini. Dan lagi, kita tidak tahu apkah kegagalan kandidat seharusnya dari kurangnya pemahaman terhadap teks atau terhadap pertanyaan. Seorang kandidat mungkin menjawab beanar sebuah soal dengan menyingkirkan jawaban-jawaban yang salah, suatu keterampilan berbeda dari kemampuan memilih jawaban yang benar pada tempat pertama.
2. Skor-skor yang dicapai dalam test pilihan ganda, seperti dalam test benar-salah, mungkin dicurigai karena calon telah menerka semua atu bebrapa jawaban. Ini berakibat penyempitan skor. Format dari test-test ini mendorong kandidat untuk menebak semuanya, kadang-kadang perlu dipertimbangkan untuk mengambil langkah mengecilkan hati kandidat untuk melakukan itu. Mungkin juga untuk menyelesaikan soal tanpa melihat teks yang ada, jika ini terjadi, apapun yang diujikan tidak bisa menjadi pemahaman teks.
3. Test-test pilihan ganda lebih lama dan lebih mahal dan lebih sulit untuk mempersiapkannya daripada beberapa ujian open-ended seperti komposisi. Banya soal yanmg ahrus ditulis dengan hati-hati oleh penulis soal yang terlatih dan meskipun telah diuji coba sebelum digunakan pada ujian format. Tiap soal harus di edit dengan teliti untuk memastikan bahwa:

- Tidak ada informasi yang berguna.
- Ejaan, grammer (tatabahasa) dan tanda bacanya benar.
- Bahasanya singkat dan sesuai untuk kandidat.
- Informasi yang cukup untuk menjawab pertanyaan.
- Hanya ada jawaban yang benar.
- Pengganggu adalah kesalah tetapi masuk akal dan mendiskriminasikan pas level yang benar.
- Pilihannya homogen, panjangnya sama, tidak ada sangkut pautnya, dan soal itu sesuai untuk test.

4. Menghasilakn banyak waktu dan persyaratan untuk memperoleh sejumlah soal yang memuaskan yang diharuskan untuk bagian. Khususnya untuk menguji keterampilan seperti berselancar. Masalah yang khusu terletak pada penentuan jawaban pengecoh yang sesuai bagi soal-soal yang menguji keterampilan reseptif yang lebih luas. Heuston (1975) mencatat bahwa lebih berguna membuat pertanyaan-pertanyaan open-ended daripada soal-soal pilihan ganda; sebaliknya para siiswa harus mengingat 4 atau 5 pilihan bagi tiap soal sementara mereka mencoba untuk mengolah teks.
5. Keberatan berikutnya terhadap penggunaan pilihan ganda yaitu bahaya dari format yang mempunyai efek yang tidak seharusnya terhadap penilaian sifat. Ini sudah menjadi bukti bagi hubungan yang rendah dengan penilaian membaca alternatif dan dengan data validitas eksternan concurren lainnya pada kemampuan membaca para kandidat (lihat Weir, 1983a).
6. Ada keraguan yang benar tentang validitas mereka seperti penilaian kemampuan bahasa; menjawab soal pilihan ganda merupakan sebuah tugas yang tidak nyata, seperti dalam kehidupan nyata, seseorang jarang dihadapkan dengen empat pilihan untuk membuat pilihan pemahaman yang baik.

Normalnya, ketika diperlukan suatu pemahaman terhadap apa yang telah dibaca atau didengar bisa disampaikan lewat ucapan atau tulisan. Dalam test pilihan ganda, pengganggu memberikan pilihan untuk berpikir, tapi mungkin malah sebaliknya memberikan kesempatan untuk tidak berpikir. Jika ada

pandangan yang berbeda mungkin dibantah bahwa kadang-kadang ada lebih dari satu jawaban yang benar untuk beberapa pertanyaan, khusus pada tingkat akhir. Apa yang penyusun test duga sebagai jawaban yang benarmungkin tidak bagi apa yang lainnya, yaitu harus jelas dalam sebuah teks.

5.1.2 Pertanyaan-pertanyaan Jawaban Singkat

Ini merupakan pertanyaan-pertanyaan di mana para kandidat harus mencatat jawaban yang spesifik pada tempat yang disediakan dalam lembar jawaban. Tknik ini benar-benra berguna untuk ujian pemahaman membaca dan menyimak. Pendapat-pendapat di bawah ini berkenaan dengan membaca juga dapat dipakai pada ujian menyimak.

Manfaat

1. Jawaban tidak disediakan seperti pada pilihan ganda; oleh karena itu jika seorang siswa menjawab benar, itu tidak akan terjadi kecuali ia memahami teks.
2. Dengan rumusan pertanyaan yang teliti, jawaban seorang kandidat bisa singkat dan oleh karena itu sebagian besar pertanyaan mungkin disusun dalam format ini.
3. Jika beberapa jawaban yang dapt diterima dibatasi, memungkinkan untuk memberikan instruksi yang benar kepada para penguji yang menilainya.
4. Aktivitas-aktivitas seperti kesimpulan, pengenalan serangkaian, perbandingan dan penentuan ide pokok suatu teks, ini bisa dilakukan dengan efektif melalui pertanyaan-pertanyaan jawaban singkat dimana jawaban harus dicari dari teks yang tesedia.
5. Sebuah kasus yang kuat bisa dibuat konteks yang sesuai, contohnya pada test EAP, untuk penggunaan teks yang panjang dengan format jawaban singkat dibanding yang lebih representatif pada bacaan yang diperlukan dalam situasi target. Mereka juga bisa memberikan data yang lebih dapat dipercaya tentang kemampuan membaca seorang calon (lihat Engineer, 1977, untuk bukti

tentang kemampuan yang meningkat hasil dari penggunaan teks yang panjang dan Appendix I contoh dari pendekatan ini pada test TEEP).

Kerugian

1. *Kerugian* utama pada teknik ini adalah bahwa ia menyebabkan kandidat menulis dan ada beberapa soal yang mengganggu penilaian konsep yang diharapkan.
2. Perhatian diperlukan dalam penyusunan soal-soal untuk membatasi jawaban-jawaban yang mungkin dapat diterima dan banyak tulisan yang diperlukan. Dalam kasus-kasus tersebut dimana terdapat banyak pembahsan tentang hal yang dapt diterima sebuah jawaban, contohnya, pada pertanyaan-pertanyaan yang memerlukan keterampilan-keterampilan menarik kesimpulan, ada sebuah kemungkinan di mana ketidaktepatan jawaban mungkin menuntun pada ketidak-dapat dipercaya penilai. Bagaimanapun juga, sikap tidak berlebihan dan standarisasi yang teliti dari penguji akan memabntu mengurangi ini.

5.1.3 *Cloze*

Dalam prosedur *cloze*, kata-kata dihilangkan dari sebuah teks setelah pengenalan kalimat. Dasar penghilangan disusun dengan mesin, biasanya antara kata kel-5 dan ke-7. Para peserta harus mengisi tiap yang dikosongkan/celah dengan mengisi kata yang mereka pikir telah dihilangkan. Penelitian Alderson (1978a) membuktikan bahwa teks ang lebih sulit merupakan langkah yang lebih baik bagi keterampilan yang lebih rendah daripada teks yang mudah. Dia menemukan prosedur yang dapat diterima secara semantik akan menjadi yang lebih unggul dari yang lain.

Dalam perbandingan antara *cloze* dan pilihan ganda, Enginer (1977) memutuskan bahwa teknik-teknik ini mengukur aspek-aspek yang berbeda dari aktivitas membaca artinya bahwa *cloze* mengukur ‘proses’ membaca, yakni kemampuan pembaca untuk memahami teks ketika ia membacanya; sebaliknya

pilihan ganda mengukur hasil membaca-artinya kemampuan pembaca dalam menerjemahkan informasi yang abstrak terhadp nilai maknanya.

Ada satu kesepakatan yang baik dari bukti yang supportif untuk menggunakan bentuk *cloze*. (Klein-Braley, 1981, h.229) berpendapat bahwa: 'sekarang ini, hasil penelitian menggunakan test *cloze* benar-benar memberi harapan. Mereka menunjukan validitas yang tinggi, reliabilitas, objektivitas, diskriminasi, dan lain-lain yang tinggi'. Dia mengutip J.D. Brown (1979), ' sebagaimana ditunjukan dalam studi ini dan studi lainnya, ia bisa menjadi test yang valid dan relabel pada kemampuan bahasa yang kedua.'

Alderson (1978a, h.2) menggambarkan bagaimana: 'dekade terakhir, memperlihatkan peningkatan penggunaan *cloze* pada bukan pengguna asli bahasa Inggris untuk mengukur tidak hanya kemampuan pemahaman membaca tetapi juga kemampuan linguistik umum mereka dalam bahasa Inggris yang merupakan bahasa Asing.' Dia menambahkan (h.39):

> "konsensus umum terhadap studi-studi dengan prosedur *cloze* pada 20 tahun terakhir telah menjadi ukuran yang dapat dipercaya dan valid bagi pemahaman membaca, bagi para pengguna asli bahasa Inggris…. Sebagaimana tindakan pemahaman teks, *cloze* telah ditunjukan untuk membuat hubungan baik dengan jenis-jenis test lain dengan teks yang sama dan juga dengan ujian pemahaman membaca yang dibakukan."

Dia menjelaskan bahwa meskipun bukti ini tidak bagi non-*native speaker* 'tampaknya prosedur *cloze* merupakan langkah yang menarik bagi kemampuan bahasa bagi non-*native speaker*.'

Istilah *cloze* diperkenalkan pertama kali oleh W.L. Taylor (1953) yang mengambilnya dari konsep umum dari *closure* yang mengarah pada kecenderungan individu untuk mnyesuaikan pola di mana mereka telah memegang arti seluruhnya. Taylor (h.416) menggambarkan sebagi berikut: 'suatu unit *cloze* mungkin didefinisikan sebagai: beberapa peristiwa dari usaha untuk meniru keakuratan pada bagian yang dihilangkan dari sebuah "pesan" (beberapa produk bahasa), dengan menentukan bagian apa yang hilang itu'. Pembaca memahami kalimat yang terpotong seluruhnya dan menyelesaikannya. Alderson (1978, h.8) menjelaskan bahwa 'prosedur *cloze* menjadi ukuran kesamaan antara

pola-pola yang diantisipasi oleh ahli sandi yang digunakan oleh penulis dalam sandi'.

Taylor pertama kali menggunakan prosedur ini utuk mengukur sejauh mana suatu teks dapat dibaca kemudian itu diharapkan bisa mengukur ujian pemahaman membaca dan kukan mengukur seluruh kemampuan bahasa. Bagi Bormuth (1962, h.134) 'test *cloze* merupakan pengukuran-pengukuran yang sama-sama valid atas kemampuan pemahaman membaca.' Heaton (1975, h.22) berpikir bahwa: 'test *cloze* mengukur kemampuan pembaca untuk membaca sandi pada pesan-pesan yang terpecahkan dengan membuat penggantian-penggantian yang paling dapat diterima dari semua petunjuk kontekstual yang ada.'

Engineer (1977) menemukan bahwa test *cloze* yang diberikan di bawah kondisi waktu yang diatur. Memberikan indeks yang benar dan dapat dipercaya pada kemampuan para siswa. Jika dua kondisi terpenuhi: pertama, materi tekstual yang digunakan memiliki tingkat kesulitan yang sesuai dengan populasi, dan kedua, berisi sejumlah soal yang dihilangkan.

Manfaat

1. Test *cloze* mudah dalam menyusunnya dan menilainya. Jika prosedur yang menilai kata yang tepat digunakan. Mereka dituntut untuk menjadi indikator yang valid bagi seluruh kemampuan bahasa (lihat. Bormuth, 1962; Brown, 1979; Engineer, 1977; dan Oller, 1979).
2. Dengan penghilangan kata kelima pada sejumlah besar soal bisa disusun pada teks yang relatif pendek dan dapat menunjukkan tingkat konsistensi internal yang tinggi. Konsistensi ini mungkin berubah-rubah, mekipun, terikat pada teks yang dipilih, memulai penghilangan dan dasar penghilangan yang dipakai.
3. Dalam literatur test-test *cloze* sering dianggap sebagai ukuran-ukuran yang benar dan sama pemahaman membaca.

Kerugian

1. Meskipun argumen-argumen yang dikemukakan untuk kepentingan prosedur *cloze*, beberapa keraguan telah terungkap. Sebagian besar mengenai validitasnya sebagi perlengkapan ujian.
2. Alderson (1978, h.392) mengemukakan bahwa:

   Prosedur *cloze* bukan prosedur kesatuan, sejak ada satu kekurangan yang dinilai dari sifat yang bisa dibandingkan diantara test mungkin digunakan. Fakta membuktikan dengan jelas bahwa test-test *cloze* yang berbeda, dihasilkan oleh beberapa variasi khususnya dari faktor yang tidak tetap, memberikan ukuran-ukuran yang berbeda yang tidak bisa diperkirakan, khususnya pada kemampuan berbahasa Inggris sebagai bahasa Asing.

   Jika seseorang mengubah teks, mengubah dasar penghilangan, mulai dari temapt yang berbeda atau mengubah prosedur penilaian, sesorang yang mendapatkan test yang berbeda mengenai koefisien reliabilitas dan validitas dan seluruh kesulitan test.
3. Bukti ini bertentangan dengan cara membedakan metode-metode penilaian yang akan dipakai dalam menilai prosedur *cloze*. Klein-Braley (1985) memberi kesan bahwa test *cloze* merupakan pengukuran yang sangat kurang baik dengan ukuran-ukuran kemampuan umum lain yang ditentukan ketika digunakan pada satu bahasa seperti pada berbagai bahasa. Dan itu tampak bahwa *cloze* tidak sesuai dengan kelompok yang terbatas (Klein-Braley, 1985); hubungan yang lemah telah ditemukan antara *cloze* dan pertimbangan guru (Klein-Braley, 1981; 1985); *cloze* tidak tampak hubungan baik dengan test-test produktif berbicara dan menulis dan skor-skor pada *cloze* tidak bisa dihubungkan dengan *native speaker* dengan mudah sejakperforma *native speaker* berubah dari suatu test *cloze* ke test yang lain (Alderson, 1978a).
4. Prosedur *cloze* terlihat menghasilakn test-test yang lebih berhasil bagi sintaksis dan leksik dan level kalimat daripada bagi pemahaman membaca pada kemampuan umum atau inferensial atau deduktif, apa yang mungkin disebut dengan kemampaun yang lebih layak (lihat Darnell, 1968). Ini akan terlihat sesuai dengan pendapat Alderson (1978, h.99) bahwa:

"pada dasarnya *cloze* merupakan loncatan kalimat… jelasnya, fakta bahwa prosedur *cloze* menghilangkan beberapa kata daripada prase atau klausa harus membatasi kemampuannya akan pemahaman test lebih banyak daripada terhadap lingkungan dekatnya, sejak kata-kata individu tidak mempengaruhi kepaduan tekstual dan hubungan percakapan (dengan pengecualian yang nyata dan alat-alat yang bersatu padu seperti anaphora, leksikal, pengulangan, dan penghubung yang logis)."

5. Mungkin syarat yang paling penting yaiut pertanyaan tentang performa apa yang ada dalam test *cloze* yang benar memberi tahu kita tentang kemampuan bahasa peserta. Sulit untuk menafsirkan skor test *cloze* ke dalam deskripsi tentang apa yang bisa dilakukan oleh peserta atau apa yang tidak bisa dilakukan dalam kehidupan nyata.

5.1.4 Mengisi Celah Penghilangan Selektif

Mengingat penemuan-penemuan negatif baru-baru ini mengenai penghilangan *cloze*, pemilihan soal-soal untuk penghilangan berdasarkan apa yang diketahui mengenai bahasa, mengenai kesulitan dalam memahami teks, yang diketahui cara bahasam mengenai kesulitan dalam memahami teks, mengenai cara bahasa bekerja dalam teks-teks tententu. Pertimbangan linguistik digunakan untuk melakukan penghilangan dan maka lebih mudah untuk meneruskan apa yang diharapkan tiap test untuk mengukur (lihat Alderson, 1987a, h.397; Klein-Braley, 1981, h.244; dan Weir, 1983a). teknik ini disebut sebagai mengisi celah penghilangan selektif bukan *cloze*.

Manfaat

1. Penghilangan selektif memungkinkan pembina test untuk memutuskan di mana penghilangan dibuat dan memutuskan pada soal-soal yang telah diseleksi berdasarkan teori dan menjadi penting bagi para peserta target tertentu.
2. Juga mudah bagi penulis test untuk membuat beberapa perubahan menunjukan untuk menajadi analisis soal yang penting dan mempertahankan

beberapa soal yang dikehendaki. Ia mungkin melibatkan penghapusan soal yang be;um memuaskan perihal membedakan nilai kecakapan.

Kerugian

1. Ini penting untuk menekankan bahwa tehnik ini membatasi satu untuk penarikan contoh pada keterampilan-keterampilan yang memungkinkan (yakni kemampuan-kemampuan yamg secara kolektif enggambarkan keterampilan membaca) daripada melaksanakan format jawaban singkat atau pilihan ganda (lihat Weir). Sedangkan pertanyaan-pertanyaan jawaban singkat dan pilihan ganda membolehkan penarikan contoh keterampilan-keterampilan membaca lebih banyak, mengisi celah lebih terbatas dimana hnaya satu kata yang dihilangkan.
2. Jika tujuan suatu test untuk mencoba keterampilan yang memungkinkan termasuk keterampilan yang lebihluas seperti berselancar, maka format tamahan intik ‘mencari celah’ diperlukan sekali.

5.1.5 C-Test

Baru-baru ini sebuah alternatif bagi *cloze* dan mencari celah penghilangan selektif telah muncul untuk test pemahaman pada unsur-unsur linguistik yang lebih spesifik dalam sebuah teks. Penyesuaian teknik *cloze* yang disebut C-Test yang telah dikembangkan di jerman oleh Klei-Braley (1981, 1985; Klein-Braley dan Raatz, 1984) yang berdasarkan pada dasar pemikiran teoritis yang sama seperti *cloze*, viz., yang menguji kemampuan untuk mengatasi pleonasme (kelebihan) yang direduksi dan membuat ramalan dari konteks.

Dalam C-Test setiap kata kedua di dalam sebuah teks dihilangkan sebagian. Dalam usaha untuk memastikan solusi para siswa diberi setengah dari kata pertama yang dihilangkan. Peserta ujian menyelesaikan kata di atas kertas ujian dan prosedur penilaian satu kata yang tepat dilakukan.

Manfaat

1. Dengan menggunakan C-Test variasi teks dianjurkan, dan diberikan soal dalam jumlah besar yang bisa disebabkan oleh teks yang sederhana, selanjutnya ini mempertinggi sifat dasar yang refresentatif dari bahasa yang sedang dicoba. Nomalnya minimal 100 penghilangan yang dibuat dan lebih representatis pada sebagian daripada yang mungkin di bawah teknik *cloze*.
2. Tugas bisa dinilai secara objektif, karena jarang ada lebih dari satu jawaban yang memungkinkan dari satu celah.
3. Sedangkan dalam *cloze* performa *native speaker* pada test merupakan faktor yang tidak tetap, menurut Klein-Braley (1985) lebih lazim bagi *native speaker* untuk bisa mendapatkan skor 100% pada C-Test. Mungkin karena beberapa bantuan dalam mengurangi skor, contohnya apa keuntungan mendapat nilai.
4. C-Test hemat dan hasil yang diperoleh sekarang mendorong reliabilitas dan validitas internal dan ekternal. Akan terlihat menggambarkan alternatif yang aktif pada prosedur *cloze* dan *selektive deletion gap filling.*

Kerugian

1. Memperlihatkan secara relatif teknik dalam bentuk ini terdapat bukti yang sedikit empiris dari nilainya. Sebagian besar perhatian telah diberikan mengenai yang dapat diterima umum sebagai ukuran kemampuan bahasa. Menari mengetahui bahwa Davies (1965) memiliki versi dari teknik ini dalam rentetannya di mana tulisan pertama satu kata diberikan.
2. Teknik ini cacat dari fakta bahwa ia menjengkelkan bagi para siswa yang harus mengolah teks yang rusak dengan keras dan validitas luar prosedur rendah.

4.1.6 *Cloze Elide*

Sebuah teknik yang menarik yaitu di mana kata-kata yang tidak semestinya disisipkan ke dalam bacaan dan para peserta harus menunjukkan di mana letak sisipan tersebut. Kenyataanya, tidak ada yang baru mengenai teknik

ini, Davies menggunaknnya baru-baru ini (Davies, 1985). Dalam bentuknya yang lebih baru, ia dikenal sebagai teknik ata-kata yang kacau.

Manfaat

1. Pada perbandingan dengan format pilihan ganda atau pertanyaan jawaban singkat, peserta tidak bermasalah dalam memahami pertanyaan. Kira-kira ia mempunyai persamaan, ia disebut sebagai test *cloze*.

Kerugian

1. Penilaiannya bermasalah karena peserta mungkin mencoret soal-soal yang benar, tetapi berlebih-lebihan.

4.1.7 Transfer Informasi

Dalam ujian baik pemahaman membaca maupun menyimak kita telah dihubungkan dengan masalah pengukuran yang menajdi '*muddied'* dengan harus menggunakan tulisan terhadap jawaban catatan. Usaha untuk menghindari kontaminasi pada skor-skor beberapa papan ujian di Inggris telah memasukkan tugas di mana informasi yang dikirimkan secara verbal ditransfer menjadi bentuk non-verbal, contohnya dengan mengisi diagram, menyelesaikan grafik, atau mengurutkan peristiwa (lihat Appendix V untuk contoh yang menarik dalam test JMB).

Manfaat

1. Teknik transfer informasi cocok untuk ujian pemahaman proses, klasifikasi atau urutan narasi dan berguna untuk ujian variasi jenis teks yang lain. Ia menghindari kemungkinan kontaminasi dari para siswa yang harus mengisi penuh jawaban.
2. Itu merupakan tugas yang realistis untuk situasi-situasi yang berbeda dan minatnya dan keaslian memberikannya validitas luar dalam konteks ini.

Kerugian

1. Banyak sekali perhatian yang harus diberikan tugas non-verbal itu peristiwa harus menyelesaikan proses. Pada beberapa tugas para siswa mungkin harus memahami teks.
2. Ada bahaya bias budaya dan pendidikan. Pada subjek tertentu para siswa mungkin juga diragukan, contohnya, beberapa siswa kelas sosial mungkin tidak mahir dalam mengerjakannya dalam medium non-verbal seperti teman-temannya di kelas IPA.

Kesimpulan

Untuk ujian kemampuan membaca kami akan menganjurkan penggunaan soal-soal jawaban singkat dan mengisi celah hilang selektif secara bersamaan. C-Test merupakan alternatif yang menarik berikutnya dan hal yang dapat dipercaya para siswa yang validitasnya bergunak bagi investigasi selanjutnya. Jika kita harus mengembangkan sifat dasar yang komunikatif dari test-test kita, mungkin penting untuk memusatkan pada tugas-tugas performa di dalam test-test membaca, dan penggunaan teknik-teknik transfer informasi dan format-format respon terbatas yang dianjurkan.

5.2 Ujian Menyimak Pemahaman

5.2.1 Ujian Kemampuan Menyimak Secara Ekstensif

Dasar pemikiran di belakang konstruksi beberapa test pemahaman menyimak sekarang ini dijelaskan oleh Valette (1967, h.49): ‘objek pokok test menyimak adalah evaluasi pmahaman siswa. Tingkat pemahamannya akan tergantung pada kemampuannya membedakan fonem, mengenal tekanan dan intonasi, dan memahami apa yang ia dengar.’

Terpikir bahwajika seorang pelajar diuji pada perbedaan fonem, tekanan dan intonasi, jumlah sub-test yang ‘berbeda’ akan sama dengan kemampuannya dalam menyimak. Contoh test jenis ini adalah rangkaian test ELBA yang dikonsep oleh Ingram (1964) yang menekankan pada soal-soal menyimak yang ‘berbeda’ seperti pengenalan suara, intonasi, dan tekanan, penggunaan soal-soal singkat

daripada penggunaan percakapn atau dialog yang terus-menerus. Sebagaimana ryan (1979) jelaskan, bahkan bagian yang disebut sebagai pemahaman menyimak terlihat lebih banyak merupakan test mekanisme respons yang sesuai dengan test pemahaman percakapan yang terus menerus pada konteks asli.

Kecenderungan yang nyata di tahun-tahun ini adalah usaha membedakan antara test-test diskriminasi yang berhubungan dengan pendekatan dan test-test yang berhubungan dengan konteks pemahaman menyimak. Templeton (1973) menjelaskan bagaimana penelitian mulai memusatkan pada test-test integratif. Pemahaman menyimak ini yang merujuk pada test-test poin berbeda pada diskriminasi fonem, intonasi dan kata, dan tekanan kalimat.

Sejak 1969 JMB bukan lagi merupakan test-test keterampilan yang berhubungan dengan pendengaran individu yang terisolasi, tetapi merupakan test pemahaman menyimak pada gabungan konteks ceramah atau dialog (lihat McEldowney, 1976, dan Appdedix IV). Perubahan paradigma ini juga bisa diobservasi dengan versi EPTB tahun 1977 (lihat Davies, 1978) yang menggantikan tugas-tugas analisis, diskriminasi fonem, tekanan, dan intonasi, seluruh muatan sub-test pemahaman menyimak, contohnya, test gabungan dari pemahaman menyimak yang berdasarkan pada ceramah dengan mengambil catatan yang disimulasikan.

Davies (1978, h.16-8) menjelaskan bagaimana perubahan-perubahan yang sama yanga terjadi antaratugas-tugas menyimak yang dijelaskan dalam buku valette (1976) dan Valette (1977) sebagai perubahan dari linguistik ke sosiolinguistik, dari strukturalisme menuju fungtualisme, dari taksonomi dan perincian ke dalam keterampilan, kedalam bagian-bagian yang berbeda, integrasi dan penambahan menjadi penyeluruh. ‘Dalam vallete edisi kedua (1977) Davies berpendapat (h.147): ‘perubahan dari memusatkan pada suara, hasilpercakapan, fonologi, kedalam makna dan komunikasi.’

Argumen kuat yang menentang diskriminasi yang berhubungan dengan pendengaran sebagai test kemampuan pada pemahamna menyimak adalah bahwa kemampuan untuk membedakan antara fonem-fonem meskipun tidak menyatakan kemampuan untuk memahami pesan verbal. Bagi Valette (1977, h.102):

‘perhatian khusus penguji yaitu untuk mengetahui apakah para siswa menerima pesan yang dimaksud dan bukan pada pembuatan diskriminasi suara tertentu atau identifikasi ciri-ciri struktural tertentu.

J.W. Morrison (1974) setelah menilai pemahaman menyimak, ia menyimpulkan bahwa dalam test ESP performa komunikatif harus diertimbangkan pada level yang melebihi test fonologi dan gramatikal, kemudian membuat laporan mengenai konteks komunikatif percakapan. Chaplen (1970, h.19) baru-baru ini menyimpulkan bahwa: ‘apapun kontribusi dari unsur-unsur komunikasi-intonasi, tekanan, diskriminasi fonemik-terhadap test komunikasi, kepenringan mereka terlihat paling rendah pada beberapa tingkat kecakapan di tingkat dasar.’

Holes (1972) mengembangkan instrumen test yang memffokuskan pada kemampuan mengenai ceramah-ceramahnya.

a. Soal-soal Pilihan Ganda

Pada pertimbangan kita mengenai penggunaan teknik ini dalam penilaianyang telah dibahas pada 4.1.1, jelas bahwa *Kerugian* menggunakan test ini diikuti oleh beberapa *Manfaat* yang mungkin ia miliki.

Karena masalah-masalah yang berhubungan dengan sifat dasar utama dari proses menyimak menyebabkan adanya kesulitan-kesulitan tambahan dalam menggunakan tehnik ini seperti ukuran kemampuan menyimak, contohnya beban tambahan yang diberikan pada pengolahan dengan harus mengingat emapat pilihan (Heaton, 1975). Formatnya dibuat dan makin terasa sebagai metode yang invalid untuk menilai pemahaman oleh para guru, para perencanaan materi dan para penguji bahasa. Sertifikat umum yang baru dari ujian pendidikan menengah (GCSE) di Inggris tidak akan menggunakan format pilihan ganda secara besar dikarenakan komentar pedas dari organisasi guru atas valditasnya sebagai pengajaran dan ujian.

Ujian RSCA CUEFL, meskipun upaya-upaya untuk memaksimalkan keaslian pada teks pendukung diseleksi, benar-benar tergantung pada format ini dan perbedaannya (contohnya, soal-soal benar-salah) dan telah dikritikkarena

kemunduranya dari pengolahan percakapan realistis yang terlibat (lihat Appendix III). Penggunaanya oleh RSA atas format-format yang lebih objektif ini menyoroti perlunya mencoba menentukan realisme stimulus teks dan tugas yang diharapkan bisa diberikan pada siswa, dan kadang-kadang tidak sesuai dengan reliabilitasnya dan validitas.

2. Soal-soal Jawaban Singkat (SAQs)

Manfaat

1. Soal-soal jawaban singkat bisa menjadi aktivitas yang realistis bagi ujian pemahaman menyimak, contohnya, jika seseorang berharap menyimulasikan aktivitas-aktivitas kehidupan nyata seseorang menyampaikan pesannya. Dengan perhatian yang cukup, jawaban bisa dibatasi dan kemudian bahaya proses 'menulis' yang mengganggu pengukuan menyimak sebagian besar dihindari (lihat Appendix I).
2. Berbeda dengan format-format pilihan ganda dan benar salah yang digunakan beberapa ujian, seseorang bisa menjadi lebih khusus yaitu bahwa jawaban-jawaban yang benar tidak datang secara kebetulan.

Kerugian

1. Jika peserta harus menulis jawaban pada waktu yang sama seperti menyimak percakapan yang terus-menerus maka terdapat masalah-masalah yang nyata. Muatan yang tidak penting mungkin dimasukan kedalam ingatan dan informasi yang sangakt penting dari percakapan yang terus menerus mungkin hilang ketika jawaban soal sebelumnya sedang ditulis.

c. Teknik Transfer Informasi

Teknik ini sudah dibahas di atas yang berhubungan dengan membaca dan berguna untuk pertimbangan alasan-alasan yang sama pada test menyimak (lihat Appendix V).

Manfaat

1. Manfaat utama menggunakan teknik ini dalam ujian menyimak adalah bahwa siswa tidak harus mengolah soal-soal tertulis ketika mencoba pemakaian yang diucapkan yang masuk akal. Ini efisien khususnya untuk ujian sebuah pemahaman rangkaian proses yang berhubungan dengan teks dan klasifikasi.

Kerugian

1. Sangat sulit untuk menemukan teks-teks yang diucapkan yang menuntun dirinya sendiri pada fomat non-verbal. Sedangkan dalam ‘membaca’ sejumlah editing teks dapat dikerjakan dengan mudah dan dalam variasi teks yang lebih besar lebih tersedia dengan mudah, ini bukan merupakan kasus bagi teks menyimak dari sumber-sumber autentik.

5.2.1 Batasan-batasan untuk Ujian Menyimak Ekstensif

Perlu diketahui bahwa jika sesorang berharap membuat tugas-tugas test lebih mirip dengan yang ada dalam kehidupan nyata, sifat dasar percakapan yang diperpanjang dan masalah-masalah pengolahan yang lebih besar berhubungan dengan pemahaman soal-soal yang menghalangi ucapan bahasa Inggris yang memusatkan pada keterampilan linguistik yang lebih khusus seperti menentukan mkana kata-kata dari konteks yang mengenalkan nilai makna dari ciri-ciri khusus tekanan atau intonasi. Benar-benar sulit bagi siswa untuk kembali dan memfokuskan pada ciri-ciri yang sangat spesifik dari percakapan ketika menyimak dan mencoba memahami percakapan non-interaktif, monolog yang terus menerus. Oleh karena itu, untuk mempertahnkan sifat dasar integratif test, kita harus memusatkan perhatian pada soal-soal keterampilan mengolah yang lebih umum seperti menyimpulkan, menyimak yang pokok-poko dan menentukan ide utama.

Masalah yang serius dalam ujian menyimak ekstensif dengan menggunakan tape recorder yaitu bahwa unsur visual, banyaknya referensi eksoporik biasa dan informasi paralinguistik yang lebih sulit bagi para kandidat. Biasanya pendengar tidak harus mengolah suara-suara yang tidak ada dari tape

recorder di kehidupan nyata (selain dari pengecualian yang nyata seperti mendengarkan radio).

Sampai ada hal yang dapat masuk ke dalam peralatan video, kepalsuan tugas menyimak akan tetap menjadi masalah. Bahkan video mungkin memiliki kesulitan prakteknya sendiri meskipun beberapa *screen* mengharuskan semua penontonnya diperlukan sama dalam ketidak cocokan dengan sistem-sistem yang berbeda.

Ada suatu bahaya yang besar dalam test-test menyimak di mana para peserta mungkin diharapkan untuk memecahkan kesulitan-kesulitan lain yang muncul dari konteks terbatas yang ada dan langkah-langkah selanjutnya harus diambil untuk mengimbangi ini atau harus meremehkan kemampuan mengolah bahasa lisan dengan serius.

5.2.2 Ujian Menyimak secara Intensif

Kesulitan-kesulitan dalam memusatkan poin-poin menyimak yang khusus telah dibahas di atas, dimana para peserta ditunjukkan pada percakapan yang terus menerus perlu mempertinggi reliabilitas rangkaian test kita sebaiknya memasukan beberapa soal yang spesifik. Dikte atau mengingat tes yang didengar dapt memberikan perbedaan ini sebaik menjadi valid dalam isi untuk kelompok-kelompok peserta tertentu, khususnya yang menyangkut studi akademis melalui medium bahasa Inggris.

Dikte

Hal ini sangatlah penting mengingat para peserta akan dinilai seteliti mungkin dimana mereka akan diharuskan untuk menggunakan bahasa. Bagi dikte, ini melibatkan, mereka menyimak materi yang didiktekan yang memasukkan pesan lisan yang khas mungkin mereka alami dalam situasi target.

Manfaat

1. Memperhatikan reliabilitas sebaik validitas, mungkin sebaliknya untuk memperbaiki seluruh rangkaian ‘menyimak’ dengan memasukan format yang

memiliki laporan yang jarang terbukti pada respek ini. Dikte dapat memberi reliabilitas ini melalui beberapa soal yang bisa dihasilkan sebaik dijadikan valid bagi situasi-situasi tertentu di mana dikte mungkin menonjol seperti aktifitas kelompok target.

2. Ada beberapa bukti yang menunjukan bahwa dikte berhubungan erat dengan variasi-variasi test lain, khususnya dengan test-test integratif lain seperti *cloze* dan sering digunakan sebagai ukuran kecakapan umum yang berguna. Ada beberapa bukti bahwa penggunaan skema penilaian semantik (lihat Weir, 1983a) seperti terhadap sistem kata yang tepat membantu mempertinggi hubungan dengan konsep test-test menyimak yang valid lainnya.
3. Kecaman atas dikte masa lalu berasal dari sudut pandang yang banyak sekali dipengaruhi oleh linguistik struktural yang mendukung uji unsur-unsur keterampilan bahasa yang lebih berbeda dan berharap mungkin dari kemungkinan pengukuran yang keruh. Heaton (1975) berkomentar: ‘seperti alat ujian, ia mengukur begitu bnayak ciri-ciri bahasa yang bebeda untuk menjadi efektif dalam memberikan cara untuk menilai keterampilan khusus seseorang. Bagaimanapun juga, para pendukung dikte mempertimbangkan sifat dasarnya yang sangat ‘integratif’ agar ber*Manfaat* sejak ia menggambarkan dengan tepat bagaimana orang-orang mengolah bahasa dalam konteks kehidupan nyata.
4. Minat baru pada dikte menggambarkan perubahan paradigma dalam nilai-nilai unjian dan tujuan-tujuannya telah dibahas. Padahal pada tahun 1967 Valette telah meneliti bahwa para spesialis bahasa asing tidak setuju dengan keefektipan dikte sebagai suatu ujian beberapa siswa yang lebih maju, sepuluh tahun kemudian dia bisa menguraikan bahwa dikte merupakan ukuran yang tepat bagi seluruh kecakapan dan metode yang luar biasa untuk mengelompokkan para siswa yang baru masuk berdasarkan tingkat kemampuannya.
5. Faktor yang penting kembalinya dikte pada popularitas sebagai alat ujian merupakan penelitian yang dilakukan oleh Oller, yang membentuk minat yang lebih besar terhadap ujian integratif. Oller (1979) menolak kecaman

dikte dan membuktikan bahwa ia merupakan test pemahaman menyimak yang tepat karena ia menguji rangkaian keterampilan integratif yang luas.

6. Oller mengklaim bahwa proses analisis dinamis dengan sintesis itu berbelit-belit. Dikte menarik kemampuan pelajar untuk menggunakan semua sistem bahasa bersama dengan pengetahuan tentang dunia, konteks, dan lain-lain, untuk memperkirakan apa yang dikatakan (perpaduan pesan) dan setelah pesn diucapkan diperiksa dengan teliti lewat ingatan singkat untuk mengetahui apakah ia sesuai dengan apa yang telah dipekirakan.
7. Bagi Oller, dikte menguji tidak hanya kemampuan siswa untuk membedakan unit-unit fonologikal tetapi juga kemampuannya untuk membuat keputusan mengenai batas-batas kata; dengan cara ini seseorang yang diuji menemukan urutan kata-kata dan frase-frase dan dari ini ia membangun ulang suatu pesan. Identifikasi kata-kata dari konteks seperti dari suara-suara yang diterima terlihat oleh Oller sebagai *Manfaat* dikte yang positif maka kemampuan ini sangat penting dalam membuat bahasa menjadi berfungsi. Keberhasilan membangun ulang pesan oleh siswa dikatakan tergantung pada tingkat 'tatabahasa harapan' yang diinternalkannya meniru itu dari *native speaker*. *Native speaker* yang mahir slalu mendapatskor 100% dari dikte yang dikelola dengan baik sementara pelajar non-native membuat kesalahan-kesalahan pada penghilangan, sisipan, perintah kata, inversi, dan lain-lain, menunjukkan bahwa tatabahasa yang diinternalkan tidak akurat dan tidak sempurna; mereka tidak sepenuhnya mengerti apa yang mereka dengar dan apa yang mereka tulis ulang dalam studi yang bebeda secaya bersamaan dari yang asli.
8. Menurut Oller, penelitian menunjukan bahwa hasil test dikte merupakan peramal kemampuan bahasa yang kuat daripada yang diukur oleh jenis-jenis test bahasa lain (lihat Oller, 1971; Valette, 1977).

Kerugian

1. Alderson (1978a) menyimpulkan bahwa bukti mengenai dikte tidak meyakinkan dan berguna hanya sebagai rangkaian test menyimak daripada sebagai solusi tunggal. Dia (1978a, h.365) menyatakan bahwa:

Alasannya adalah lebih banyak berhubungan dengan beberapa sub-test daripada dengan yang lainnya yang tidak terlihat menjadi fakta yang semestinya dikalim bahwa ia merupakan test integratif, tetapi karena secara esensial merupakan test keterampilan linguistik level awah. Oleh karena itu, dikte mempunyai hubungan paling baik dengan test-test *cloze*, teks-teks dan metode-metode penilaian yang mereka sendiri memperbolehkan pengukuran keterampilan-keterampilan ini.

2. Dikte akan menjadi hal sepele. Jika tidak memori siswa merupakan tantangan dan panjangnya kata-kata yang didiktekan akan tergantung pada kemampuan pendengar yang terbatas di man arekan-rekan *native speaker* bisa menanganinya.
3. Penilaian mungkin menjadi suatu masalah, jika sesorang berharap membawanya ke dalam keseriusan kesalahan atau jiak seseorang berharap menggunakan skema penilaian yang dikenal secara lebih komunikatif ketika nilai diberikan jika peserta telah memahami isi pokok pesan dan ciri-ciri redundan (pleonastis) diabaikan.
4. Jika dikte tidak diberikan dengan menggunakan tape, test akan kurang reliabel, seperti akan ada perbedaan-perbedaan pada kecepatan pengiriman teks kepada para pendengar yang berbeda.
5. Latihan bisa menjadi tidak realistis jika teks yang digunakan telah dibuat sebelumnya untuk dibacakan daripada didengarkan.

Listening Recall

Berbeda dari dikte, bukti yang ada mengenai listening recall sangat sedikit (lihat Furneaux, 1982; dan Beretta 1983 untuk penjelasan yang lengkap mengenai prosedur ini). Siswa diberi salinan dari bagian kata-kata tertentu yang telah dihilangkan (penghilangan ini diperiksa terlebih dahulu untuk memastikan bahwa mereka tidak bisa diisi hanya dengan membacanya). Kata-kata yang dihilangkan biasanya kata-kata isi yang dirasa penting untuk memahami percakapan dan tempat yang dikosongkan itu terjadi pada interval yang sering.

Para siswa diberi sedikit waktu untuk membaca teks, memperbolehkan penggerakan tatabahasa yang diharapkan. Mereka harus mengisi titik-titik, setelah mendengarkan tape recorder bagian yang komplit dua kali. Pertama kali mereka dianjurkan untuk medengarkan kemudian berusaha untuk mengisi titik-titik dalam waktu singkat. Mereka diperbolehkan untuk menulis jawaban-jawabannya. Mereka mendengarkan bagian yang kedua kali dan kemudian dalm waktu singkat untuk menulis beberapa kata yang hilang. Waktu menulis yang terbatas ini mendatangkan memori yang singkat. Format ini melibatkan beberapa faktor linguistik yang telah dibahas untuk dikte dan direfleksikan kedalam nama-nama lain yang telah diberikan seperti dikte spot, dan kombinasi *cloze* dan dikte.

Manfaat

1. Seperti dikte, ia bisa dilakukan dengan cepat dan dinilai secara objektif dan membiarkan penguji memutuskan pada soal-soal yang dianggap penting.
2. Hubungan erat dengan test menyimak yang lebih langsung lainnya (Beretta, 1983) dan dengan total test untuk rangkaian menyimak telah disampaikan.
3. Memiliki keuntungan untuk pelaksanaan ujian dalamskala besar yaitu mudah merencanakan, melaksanakan, dan menilainya.

Kerugian

1. Kesulitan menggunakan teknik ini terletak pada perumusan apa yang diujikan. Sebagmana halnya satu kata yang dihilangkan ia mungkin tidak menguji apapun lebih banyak dari kemampaun untuk menyesuaikan suara-suara dengan simbol-simbol dibentuk oleh kemampuan untuk membaca bagian yang ada celahnya.
2. Ia merupakan tugas yang tidak autentik dan melibatkan kemampuan membaca sebaik kemampuan menyimak. Konsep yang teliti diperlukan untuk memastikan bahwa para siswa tidak dapat mengisi titik-titik hanya dengan membaca bagian tanpa harus mendengarkan semuanya.
3. Memberikan korelasi yang tinggi yang telahdiketahui antara *listening recall* dan dikte (lihat Furneaux, 1982; Beretta, 1983), dan kepraktisan padan kata

dan reliabilitas, validitas potensial dikte yang lebih besar bagi kelompok-kelompok tertentu, contohnya, bagi para siswa yang belajar melalui medium bahas Inggris, mungin menuntun pada referensi untuk dikte lebih dari *listening recall.*

4. Permasalah yang mungkin terjadi dalam penilaian dengan pertimbangan diberikan kepada apapun daripada *spelling* yang tepat.

Kesimpulan

Test-test menyimak memungkinkan untuk memasukkan tugas performa yang asli. Suatu upaya akan dibuat untuk teknik-teknik transfer informasi yang tidak tergabung (lihat appendix V). kita mungkin memasukkan soal-soal jawaban singkat dengan berguna dan pertimbangan bisa diberikan kepada dikte (lihat Appendix I).

Ujian Menulis

Dua pendekatan bagi penilaian kemampuan ‘menulis’ bisa digunakan. Pertama, menulis bisa dibagi kedalam tigkat-tingkat yang berbeda, contohny, tatabahasa, kosa kata, ejaan, dan tanda baca, dan unsur-unsur ini bisa diuji secara terpisah dengan menggunakan test-test objektif. Kedua, tugas-tugas menulis yang diperluas lebih langsung dari jenis-jenis yang berbeda bisa dikonsep. Insi semua akan menjadi validitas konsep, isi, luar, dan *wasback* yang lebih besar tetapi akan membutuhkan penilaian yang lebih subjektif.

5.3.1 Metode tidak Langsung untuk Menilai Kompetensi Linguistik

Pada bagian 4.1, kita menguji format-format *cloze,* mengisi celah hilang selektif, C-Test, dan menguraikan nilai dari teknik-teknik ini untuk ujian linguitik yang lebih spesifik, keterampilan membaca kalimat, *viz,* soal-soal yang memusatkan pada pemahaman kosa kata, struktur, atau alat kepaduan.

Baik kemampuan produktif maupun ketereampilan reseptif bisa dierinci kedalam tingkatan tatabahasa dan leksikal berdasarkan kerangka poin yang

berbeda. MCEldowney (1974, h.8) menguraikan test silabus JMB pada bahasa Inggris (bahasa asing):

Untuk bisa melaksanakan keempat keterampilan ini (menyimak, membaca, berbicara, dan menulis) dengan fungsi yang berbeda, penting untuk bisa memanipulasi soal-soal dari ketiga tingkatan bahasa. Yaitu, untuk berkomunikasi harus mempunyai kosa kata yang cukup, untuk mengetahui penggunaan tatabahasa bahasa Inggris dan untuk dapa menangani suara-suara bahasa Inggris, tekanan, dan intonasi.

Test JMB dalam bahasa Inggris (bahasa asing), seperti memasukkan tugas-tugas ujian hasil tertulis juga mempunyai tugas-tugas yang menguji pengetahuan mengenai ‘kosa kata produktif dasar’ dan ‘soal-soal grammatikal’ (lihat Appedix v). permasalah yang mengahadapkan konsep test-test kosa kata bermacam-macam. Chaplen (1970an) yang membuat konsep sub test bagian kosa kata dari test-test JMB baru-baru ini mencatat dua bidang masalah utama:

1. Pemilihan soal-soal leksikal;
2. Metode-metode yang digunakan untuk menguji soal-soal leksikal.

Jika orang-orang yang diuji mempelajari variasi subjek yang berbeda, maka terdapat suatu masalah yang serius dalam pemilihan. Lebih bnayak materi yang digeneralisasi, lebih sulit untuk menyeleksi tujuan ujian. Dalm bidang-bidang yang diujikan khusus, dimana terdapat sesuatu yang dapat diidentifikasi, register yang disetujui, seleksi lebih mudah tetapi masih tetap sulit.

Permasalahan lain yang terjadi pada bobot relatif yang akan diberikan kepada soal-soal yang dipilih dari materi-materi membaca yang akan datang seperti terhadap soal-soal yang mungkin digunakan dalam tugas-tugas menulis yang luas. Apakah kita menguji kosa kata aktif atau pasif? Selanjutnya, bagaimana kita menentukan tingkat frekuensi dan pentingnya soal-soal leksikal yang diperluas untuk digunakan dalam test?

Permasalahan yang sama terjadi pada seleksi soal-soal grammatikal untuk dimasukkan kedalam test kompetensi linguistik yang tidak langsung. Penelitian kuantitatif terhadap pembuatan soal-soal struktural pada materi-materi tertulis reseptif dan produktif, populasi test harus mengatasinya pada situasi target yang

akan datang. Yang lebih pragmatis, metode pengambilan keputusan yang subjektif terhadap soal-soal yang akan diberikan itu diperlukan. Tampak perlu untuk menguji muatan test yang ada dan isi buku pelajaran pada tingkat yang sama untuk menentukan apa yang para ahli di bidangnya anggap sebagai soal-soal yang cocok bagi populasi yang sama.

Juga terlihat ada suatu masalah dalam menyampaikan apa yang sedang diujikn dalam test tatabahasa pon yang berbeda ini. Akankah performa test pengetahuan grammatikal tidak langsung disampaikan menurut profil untuk membaca dan menulis? Teknik tidak langsung lainnya terbatas pada persoalan yang berkenaan dengan validitasnya bagi para peserta ujian dan para pengguna hasil test tersebut. Usaha yang menarik untuk memelihara keobjektifan dan ulasan pendekatan poin yang berbeda ketika mempertinggi validitas bisa ditemukan pada tugas editing pada buku kedua mengenai test TEEP (lihat Weir, 1983a; dan Appendix I).

Tugas Editing

Dalam tugas editing siswa diberi teks yang berisi sejumlah kesalahan tatabahasa, ejaan, tanda baca, dari jenis-jenis yang ditulis biasa oleh para guru remedial bagi para siswa dalam kelompok target dan dimintai untuk menulis ulang bagian yang membuat semua koraksi penting.

Manfaat

1. Seperti pengukuran kompetensi yang lebih objektif, tugas ini mungkin memiliki efek washback yang baik di mana para siswa diajar dan didorong untuk memperbaikai tugas tertulis mereka dengan lebih teliti.
2. Pasti lebih banya *face valid* daripada teknik tidak langsung lainnya seperti menyesuaikan bagian proses menulis.

Kerugian

1. Jika siswa menulis ulang bagian-bagian kalimat dengan menggunalan kata-katanya sendiri daripada hanya mengoreksi kesalahan-kesalahan, maslah-

masalah penilaian menjadi pertimbangan. Juga terdapat beberapa keraguan seperti apakah kemampuan untuk mengoreksi kesalahan-kesalahan orang lain sama dengan mengoreksi kesalahan-kesalahnnya sendiri.

2. Penilaian juga bisamenjadi problematis jika peserta mengubah sesuatu yang sudah dikoreksi, sebuah kesalahan panitia bukan kelalaian.

5.3.2 Ujian Menulis Langsung

Dengan pendekatan yang lebih integratif dan langsung atas ujian menulis, kita bisa menggabungkan soal-soal yang menguji kemampuan peserta untuk mengerjakan tugas-tugas fungsional yang diperlakukan dalam performa tugas dalam situasi target. Bagi para dokter disebuah rumah sakit ini mungkin menyangkut penulisan surat kepada GP lokal mengenai seorang pasien. Bagi seorang siswa dalam konteks EAP mungkin menyangkut pencarian teks akademi untuk mengutip informasi yang spesifik yang akan digunakan pada rangkuman tertulis (lihat Appendix).

Test-test Essay

Ini merupakan sebuah metode tradisional bagi para siswa untuk membuat contoh tulisan yang berhubungan. Stimulus biasanya ditulis dan bisa mengubah panjangnya dari beberapa tulisan yang terbatas menjadi beberapa kalimat. Topik-topiknya seringkali sangat umum dan tergantung pada peserta yang mengeluarkan isi kepalanya. Para peserta bagaimanapun juga tidak dituntun seperti bagaimana mereka diharapkan untuk menjawab pertanyaan.

Manfaat

1. Secara tradisional, essay telah diakui sebagai teknik ujian yang mungin menjelaskan keengganan yang besar untuk membuangnya meskipun permasalahan dalam penilaian elah ditemukan (lihat Coffman, 1971; Gips dan Ewen, 1974).

2. Topik-topiknya benar-benar mudah dan merupkan teknik ujian yang terkenal bagik bagi para peserta maupun bagi para pengguna hasil test. Maka ia memiliki validitas luar yang dangkal bagi orang awam.
3. Merupakan sarana yang cocok untuk menguji keterampilan-keterampilan, seperti kemampuan untuk mengembangkan argumen dengan cara yang logis, yang tidak bisa diuji dengan cara lain.
4. Manfaat besar sama dengan test-test menulis yaitu bahwa penulisan contoh dibuat yang bisa memberikan poin referensi bagi perbandingan yang akan datang.

Kerugian

1. Bebas, *open-ended writing* merupakan suatu masalah. Kemampuan untuk menulis topik-topik *open-ended* yang biasa mungkin tergantung pada latar belakang dan pengetahuan kultural peserta, khayalan atau kreativitas. Ini semua mungkin bukan merupakan faktor-faktor yang kita harapkan untuk dinilai.
2. Peserta mungkin tidak berminat dengan topik yang diberikan dan jika pemilihan topik yang diberikan sangat sulit untuk membandingkan performa khususnya jika hasil dari jenis-jenis yang bebeda dimasukkan.
3. Tekanan waktu sering menjadi ketidakleluasaan yang tidak realistis bagi essay dan tidak dilakukan di luar kehidupan akademis. Bagi sebagian besar orang proses menulis itu lebih lama dan mungkin melibatkan beberapa draft sebelum versi yang sempurna dibuat.
4. Pencantuman komponen ‘menulis’ pada suatu ujian menghabiskan waktu dari total waktu test yang ada untuk menguji semua keterampilan.

Tugas-tugas Menulis yang Terkontrol

Ternyata ada kasusus yang sangat kuat untuk memasukkan test menulis kedalam bidang validitas isi dari tugas-tugas ‘contoh tugas’. Ia menguji keterampilan penting yang tidak ada bentuk penilaian yang lain bisa dibuat contoh dengan tepat. Mengabaikan tugas ‘menulis’ pada situasi dimana tugas-tugas

‘menulis’ merupakan suatu ciri penting dari kebutuhan kehidupan nyata siswa mungkin benar-benar mengurangi validitas program uji.

Wall (1982) mengadakan investigasi yang memperjelas jenis-jenis tugas ‘menulis’ siswa teknik mesin diharuskan untuk menunjukakan bagian dari tugas sekolah mereka dan membandingkannya dengan tipe-tipe essay yang mereka set dalam rangkaian Michigan yang digunakan untuk menilai kecakapn bahasa siswa untuk masuk ke universitas. Dia (h.166) membuat kesimpulan mengenai perbedaan tersebut sebagai berikut:

> Perbedaan utama terlihat bahwa dalam tugas-tugas ahli mesin ada banyak input berdasarkan teori dan tugas itu sendiri diuraikan dengan tegas, sedangkan dalam komposisi penulis hanya mempunyai harapan untuk merespons dan tidak harus hanya membuat isi tulisan tetapi juga membuat konteks, audiensi, dan tujuan dengan baik. Kriteria untuk menilai juga akan terlihat berbeda.

Kesimpulan harus investigasi itu menunggu penelitian yang menghasilkan: ‘korelasi studi antara total rangkaian Michigan dan bagian skor dengan GPA pertama siswa, yang tidak ada hubungan yang signifikan antara test dan kriteri guna kesuksesan akademis bisa ditemukan antara performa menulis dalam test dan indikator-indikator performa berikutnya dalam bida studi.

Menulis yang tidak terkontrol terlihat akan seperti test kemampuan ‘menulis’ yang tidak benar yang diperlukan oleh kebanyakan siswa. Lebih mudah untuk memperhitungkan kemungkinan dari test-test menulis ketika memperhatikan penentuan tiap tugas: media, peserta, tujuan, dan situasi, dengan aktivitas-aktivitas performa tingkat target (lihat Wall, 1982). Ketika tugas ditentukan lebih tepat dengan cara ini juga lebih mudah untuk membandingkan performa para siswa yang berbeda dan untuk memperoleh tingkat relibilitas yang lebih besar dalam penilaian. Jika tugas ‘menulis’ tidak terkontrol, para siswa yang diuji mungkin juga bisa menutupi kelemahan-kelamahannya dengan menjauhi masalah.

Ada beberapa jenis stimuli yang berbeda yang bisa digunakan pada tugas-tugas menulis yang terkontrol. Stimuli bisa ditulis, diucapkan atau non-verbal secara paling efektif, contohnya grafik, rencana tau gambar di mana siswa diminta

untuk menjelaskan dengan tulisan. (lihat apendix V; Dunlop, 1969; McEldowney, 1974, 1976, 1982;Weir, 1983a, dan Appendix I untuk contoh-contoh tersebut).

Manfaat

1. Manfaat stimuli non-verbal yaitu jika mereka memberikan informasi dengan cara yang jelas dan tepat, peserta tidak harus menghabiskan banyak waktu dalam memberi sandi suatu teks tertulis. Tugas menjadi yang paling efektif ketika peserta diminta untuk mengomentari kecenderungan khusus yang ditunjukan dalam sebuah grafik, atau untuk membandingkan dan membedakan satu set gambar dengan ambar lainnya. Stimuli yang berbeda bisa digunakan untuk memperoleh performa tertulis dari beberapa fungsi bahasa yang berbeda seperi argumentasi, deskripsi suatu proses, perbandingan dan perbedaan atau atau tulisan satu instruksi.

Kerugian

1. Permasalahn muncul ketika keinginan untuk tidak menyukai kelompok peserta tertentu manapun. Test diambil bagi daerah yang benar-benar spesial seperti jilid buku, atau pertengahan helm untuk stimulus visualnya. Para peserta sering tidak mengatasi tantangan mental untuk mengambil jenis ini dan menyerah daripada melompati lewat intelektualnya untuk mendapat tugas menulis.

   Masalah mungkin selalu terjadi ketika kerumitan stimulus menghalangi hasil yang diinginkan, yakni seseorang harus memahami instruksi-instruksi yang sangat rumit dan/atau stimuli visual untuk membuat deskripsi yang relatif terus terang mengenai proses atau klasifikasi data.
2. Kesulitan-kesulitan yang disebabkan oleh tugas-tugas tipe transfer informasi ini mungkin muncul lewat perbedaan-perbedaan yang berhubungan dengan pendidikan dan kebudayaan dalam kemampuan untuk menjelaskan grafik atau tabel atau gambar bergaris.

Ringkasan

Manfaat

1. Ringkasan bisa menjadi test yang valid, contohnya sangat tepat untuk menguji kemampuan menulis siswa dengan tugas ini. Siswa harus mengevaluasinya dalam situasi akademis. Menulis laporan atau essay membutuhkan kemampuan untuk memiliki fakta yang relevan dari banyaknya data dan untuk mengkombinasi ulang ini dalam bentuk yang dapat diterima. *Summary* yang melibatkan kemampuan untuk menulis komposisi yang terkontrol yang mengandung ide esensial 'menulis' dan membuang yang non-esensial.

Kerugian

1. Masalah dari kekhususan dari teks para peserta diharapkan membuat tulisan dalam tugas ringkasan seperti dalam tugas-tugas menulis yang terkontrol lainnya.sering dapat kesulitan dalam memilih teks stimulus yang tepat karena kekhususan subjek mereka akan menciptakan beberapa masalah bagi yang bukan ahli dan test mungkin menjadi tidak valid. Satu alternatif yaitu memilih teks yang tidak dikenal yang tidak seorangpun yang menemukan kemampuan pokok.

   Jika para siswa jurusan IPA dan teknik mesin harus membaca teks 'umum' atau teks 'netral' dan kemudian merangkum dengan menggunakan kosa kata non-sains dan menunjukkan kualitas tulisan dan imajinasinya mungkin terdapat beberapa masalah validitas yang serius bagi para siswa ini. Meskipun siswa IPA mungkin tidak bisa merangkum sebagian dan 'mengapa kucing mungkin menjadi binatang peliharaan yang cocok untuk seorang wanita tua' ia mungkin bisa merangkum keistimewaan proses yang penting.
2. Kesulitan utama pada komponen menulis yang digabungkan dari jenis ini yaitu membuat penilaian yang dapat dipercaya dan konsisten, menilai jawaban siswa dengan dapat dipercaya seseorang harus merumuskan poin-poin pokok yang masuk ke dalam kutipan, merencanakan skema penilaian yang tepat dan menstandarisasi penilaian.

Manfaat pendekatan impresionik dan analitik terhadap penilaian untuk memperbaiki reliabilitas dan valididtas sub-test ‘menulis’ akan diujikan di bawah. Perhatian sedikit diberikan kepada perbaikan reliabilitas penilaian menulis dan satu upaya telah dibuat untuk meneliti suatu bidang dan membawa bersama-sama apa yang diketahui tentang pendekatan-pendekatan pokok terhadap masalah ini. Mengenai seluruh struktur buku ini, menempati sebagian besar pembahasan metode test memberikan pentingnya keterampilan khusus yang krusial kepada para siswa yang belajar lewat medium bahasa Inggris, perlakuan yang ada dianggap ber*Manfaat*. Komentar-komentar bagi penilaian menulis menggunakan, *mutatis muntandis,* bagi penilaian hasil yang dibicarakan. Kita mempunyai hasil yang dapat diidentifikasikan yang bisa dievaluasi yang berkenaan dengan kriteria yang spesifik.

4.3.3 Penilaian Kesan Umum dan Analitik

Perbandingan dua pendekatan

Kita telah membahas bagai mana mungkin memperbaiki validitas dan reliabilitas. Kita membuat kesimpulan bahwa ada satu syarat bagi sub-test menulis yang terkontrol yaitu *register,* konteks dan jangkauan tugas menulis telah ditentukan untuk peserta. Pada bagian ini, kita menguji bagaimana aplikasi pendekatan-pendekatan yang berpengaruh dan pendekatan-pendekatan nalisis yang distandarisasi terhadap penilaian mungkin dapat membantu kita dalam upaya untuk memperbaiki reliabilitas dan validitas sub-test menulis.

Penilaian analitis menunjuk kepada metode untuk tiap kriteria yang terpisah pada skema niali diserahkan pada nilai yang terpisah dan nilai akhirnya merupakan gabungan dari penilaian-penilaian individu itu.

Metode kesan pada penilaian biasanya memerlukan dua penilai atau lebih yang memberi satu nilai berdasarkan total kesan mereka terhadap komposisi sebagai keseluruhan (lihat Luiseman, 1949; E. Ingram, 1970). Tiap kertas diberi skor dengan menggunkan skala yang disetujui dan skor seorang peserta ujian kira-kira dari nilai-nilai gabungan. Gagasan penilaian esan secara spesifiki meniadakan beberapa usaha untuk memisahkan ciri-ciri komposisi yang berbeda untuk tujuan-

tujuan penilai. Menurut Francis (1977), dalam bentuknya yang paling murni, penilaian kesan biasanya mengharuskan setiap penilai untuk membaca contoh skrip, mungkin 10-255, untuk menentukan standar dalam pikirannya dan kemudian membaca semua skrip dengan cepat dan memberi angka nilai pada setiap skrip.

Hartog dkk (1936) mengadakan satu dari beberapa studi yang ada sekarang ini kedalam keefektifan penilaian kesan analitis dan umum untuk menilai komposisi bahasa Inggris. Ada beberapa maksud untuk mengetahui metode mana yang menghasilakan hasil-hasil ytang lebih besar yang berkenaan dengan kemampuan untuk mengurangi kesalahan penilai. Penelitian ini menemukan (h.123) bahwa perbedaan antara para peneliti telah dikurangi oleh metode analitis; ‘ada beberapa ketidaksesuaian yang lebih besar antara nilai-nilai yang diketahui oleh kesan daripada anatara niali yang diketahui oleh detai terlihat bahwa ketidaksesuaian ini semestinya pada perbedaan-perbedaan yang lebih besar dari standar-standar penilaian para penguji yang berbeda ketika mereka menilai dengan menggunakan kesan’.

Penelitian juga menunjukan bahwa sebagian besar penguji menunjukan konsisten berat sebelah pada hal yang berkenaan dengan kemurahan hati dan kesederhanaan dalam penilaian mereka. Bukti yang mereka dapatkan dari ketidaksesuaian pada urutan penempatan dengan menggunakan beberapa cara lebih serius sejak pertentangan jenis ini tidak mudah bagi koreksi dengan cara yang sama seperti perbedaan yang berasal dari bias susunan nilai. Keduanya bisa dikoreksi oleh ketentuan skema nilai yang diperinci dan oleh standarisasi para penguji yang efisien berdasarkan teori untuk menilai tugas.

Seperti halnya Hartog dkk (1936), Cast (1939) juga menemukan metode analisis yang agak hebat dalam sistem penilaian tunggal. Kecamannya terhadap metode kesan yaitu bahwa menilai mereka dengan karakteristik yang lebih dangkal daripada metode analisis. Bagaimanapun juga, meskipun metode analissi mempertimbangkan yang lebih sesuai, Cast merasa bahwa hasil-hasilnya tidak memberikan bukti yang pasti dari reliabilitas peni;laian analitis yang unggul dan oleh karena itu ia menolak penggunaan salah satu metode itu sendiri.

Cast menunjukkan karakter-karakter penting yang melekat dalam dua sistem. Ciri metode analisis yang penting (h.263-4) adalah: ‘dalam merata-ratakan nilai untuk semua pertanyaan, susunannya pasti menyusut… “regresi” ini menjadi konsekuensi yang tidak dapat dihindari dari semua bentuk penyajian terakhir pada perbandingan gambar-gambar yang tidak digabungkan dengan sempurna. Dia mencatat (h.263) bahwa penilaian kesan ditentukan oleh peserta individu dan bahwa angka nilai diketahui oleh penguji yang bebeda pada skrip yang sama yang cenderung luas tidak biasanya.

Cast (h.264) juga mencatat kecenderungan penilian pesan:

> Untuk mengukur beberapa poin yang penting atau dangkal-kesalahan-kesalahan ejaan, tatabahasa atau fakta dan menyusun semua bagian: sebaliknya, metode analitis dengan menggunakan poin-poin terbatas yang banyak dan poin-poin yang tidak esensial yang memungkinkan, mungkin mengabaikan kualitas-kualitas tertentu yang menggolongkan essay sebagai keseluruhan.

Prancis (1977) juga menunjukan bahwa bahaya besar dari penilaian pesan yaitu bahwa pesan kualitas sebagai keseluruhan akan dipengaruhi oleh hanya satu aspek dari aspek-aspek kerja. Dia berpendapat bahwa prasangka dan bias dari penilaian mungkin menjadi bagian yang lebih besar dalam menentukan nilai daripada dalam skema analitis.

Penilaian Multiple (ganda)

Wiswman (1949) meneliti kemungkinan untuk memperbaiki penilaian dengan menjumlah nilai-nilai ganda dari empat penilai bebas, penilai yang tidak distandarisasi, menggunakan metode kesan yang cepat. Dia menemukan bahwa penilaian ganda dengan menggunakan metode kesan telah memperbaiki reliabilitas dan lebih cepat daripada prosedur-prosedur analitis perbandingan. Dia (h.205) memperkirakan bahwa jika rata-rata inter-korelasi sebuah kelompok dari empat penilai lebih rendah 0,6 dari yang lainnya: ‘perkiraan korelasi yang mungkin dari nilai-nilai rata-rata ‘betul adalah 0,92. Nilai ini sangat tinggi dibanding dengan yang kami perkirakan dari satu nilai analitik’.

Wiseman (h.208) mengambil contoh kesakitan menjadi stres bahwa: ‘efisiensi penilai harus menilai atas dasar konsistensinya.’ Dia berpendapat (h.2204) bahwa konsistensi efisien diperoleh dengan nilai yang murni, penilaian korelasi (penilai-intra reliabilitas), menggunakan metode penilai yang sama pada kedua kesempatan: ‘merupakan satu ukuran yang jelas atas konsistensi yang benar, dan satu yang menutup hubungan kepada konsep normal dari reliabilitas test.’ Dengan menggunakan sistem penilaian ganda yang berdasar pada prinsip konsistensi diri sendiri, dia memungkinkan untuk mencapai level reliabilitas yang tinggi.

Karya Coffman dan Kurfman (1968) dan Wood dan Wilson (1974) menggambarkan peringataan bagi permasalahan instabilitas penguji dengan menilai perilaku. Mereka telah membuat fakta bahwa menilai perilaku tidak lagi seimbang selama masa penilaian, ketika nilai yang besar dari sebuah skrip dinilai (lihat Edgewort, 1888). Mereka berargumen atas subjektifitas setiap skrip yang lebih dari satu penilai, yang mungkin membantu menetralisasi efek dari inkonsistensi penilaianm perilaku selama penilaian berlangsung.

Walau beberapa keraguan diekspresikan pada masa lampau (lihat Edgewort, 1888) mengenai kelayakan memiliki lebih daru penilai, Britton (1963), Britton dkk (1966), Lucas (1971) dan Wood Quinn (1976) semuanya menemukan bahwa penilaian ganda memperbaiki reliabilitas penilaian essay bahasa Inggris.

Britton dkk (1966), dalam sebuah eksperimen yang dirancang untku menemukan reliabel yang lain yang menilai perlengkapan untuk digunakan oleh dewan ujian, membandingkan penilaian ganda ekperimental dengan penilaian tunggal yang diangkat oleh dewan uji GCE. Mereka menemukan (h.21): ‘figur-figur itu jelas mengindikasikan bahwa dalam kasusu penilaian oleh penguji secara individu dengan uraian yang sangat hati-hati dan meneliti aransemen bagi modernisasi pada kenyataannya kurang reliable dibanding penilaian ganda. ‘ketika ofisial menilai dan penilaian ganda dikorelasikan dengan kriteria ekternal dari pekerjaan rumah yang dibuat oleh para peserta selama tahun ajaran, penilaian ganda ditmukan untuk mencocokkan.

Head (1966) membuat suatu percobaan untuk menemukan tambahan penilaian impresi daru dua penguji akan lebih reliabel dibanding dengan penguji individual. Dia menemukan (h.71): ‘kenaikan koefisien dari 0,64 bagi korelasi penilaian tunggal menjadi 0,84 untuk korelasi pasangan penilaian yang menunjukkan bahwa penilaian tambahan menjadi lebih reliabel.’

Lucas (1971) menmukan bahwa meskipun penggunaan penilai yang inkonsisten (artinya nilai korelasi hanya 0,65) dalam penilaian ganda dengan impresi menambah reliabilitas pemberian nilai yang signifikan. Pertambahan yang besar dalam reliabilitas terjadi dalam perubahan dari satu menjadi dua peniali.

Wood dan Quinn (1976) menggunakan level ‘O’ essay bahasa Inggris dan pertanyaan *summary* menemukan bahwa penilaian impresi oleh sepasang penilai menjadi lebih reliabel dibanding dengan penilai tunggal. Mereka memberi kesan, walaupun, tidak terdapat suatu keuntungan dalam reliabilitas dari suatu penilaian analitik dibanding dengan penilaian impresi tunggal. Pembaharuan yang ril adalah dalam hal penilaian.

Penilaian Holistik

Dalam evaluasi holistik, penilai berdasar kepada penilaian mereka dalam impresi mereka dari semua komposisi: dalam penilaian frekuensi (lihat Steel dan Talman, 1936), total penilai atau menjumlahkan berbagai elemen dalam komposisis, seperti: perlengkapan kohesif, kesalahan ucap kata, kesalahan peletakan koma, atau kesalahan kalimat. Jacobs dkk berpendapat bahwa metode ayang akan datang lebih objektif dan juga lebih reliabel. Validitasnya tidak begitu pasti karena satu komposisi dievaluasi oleh metode *frequency-count* yang telah dinilai bukan untuk efek komunikatif, tetapi untuk nomor atau salah satu dari elemen.

Evaluasi holistik terlihat lebih objektif dibanding impresi yang dibuat oleh para penilai. Jacobs dkk (h.29) berpendapat bahwa:

> Atas kedengkian (atau malah karena) subjektifitas ini, evaluasi holistik telah memperlihatkan kapabilitas dalam membuat penialian reliabel yang tinggi. Kebanyakan para pelajar menguji… pada faktanya, berdasar pada evaluasi holistik dari satu tipe atau yang lain dan semua pelajar

> memperoleh reliabilitas pembaca dalam tengah-tengah-menuji-tinggi delapan puluh atau sembilan puluhan. Secara intuitif hal itu terlihat bahwa komposisi skor berdasar pada respon holistik dari para pembaca yang menyertai pesan penulis harus lebih valid daripada yang berdasar pada metode *frequency-count,* yang pada pembayaran terbaik hanya sebuah kecupan bagi pendapat dan ide penulis. Sebagaiman Cooper (1977) menaruh hal itu, 'evaluasi holistik oleh responden menusia menjadikan kita lebih dekat kepada sesuatu yang esensial menusia komunikasi dibanding apa yang dilakukan oleh *frequency-count.'*

Evaluasi holistik jelas dilebih-lebihkan di man perhatian utama adalah dengan mengevaluasi ketidak-efektifan komunikatif tulisan peserta. Itu adalah sebuah kasus dalam proyek TEEP (lihat Weir, 1983a, dan Appendix I) di mana preferensi diperuntukkan bagi analitik, skema penilaian holistik dalam impresionisti yang satu, menyokong sesuatu yang eksplisit daripada daftar yangimplisit dari sautu roman atau kualitas untuk membantu para penilai.

Hal itu terasa sangat kuat, bahwa perhatian yang sangat kecil telah terbayar pada masa lalu dengan kriteria aktual yang diaplikasikan, secara implisit atau eksplisit, menjadi contoh dari pembuatan karya tulis. Sejajar dengan skema anlaitik yang diserahkan kepada para pelajar, terlalu banyak ruang bagi interpretasi idiosinkratik dari segala standar konstitut yang masih diaplikasikan kedalam skrip. Aplikasi yang bersih, kriteria yang tepat telah dirasa begitu penting.

Jacobs dkk membuat perbedaan anatara penilain holistik dan penilaian jumlah frekuensi seperti terhadap divisi yang saling melengkapi kedalam penilaian kesan dan penilaian analitis yang digunakan oleh para peneliti. Mereka menggambarkan divisi sebagai berikut: "istilah holistik Cooper berarti beberapa prosedur yang berarti "mengurutkan ciri-ciri retorikat dan ciri-ciri informasional."

Chaplen (1970an) berpendapat bahwa hasil-hasil yang lebih reliabel mungkin bisa diperoleh dari metode kesan untuk menilai apakah skala yang digunakan merupakan salah satu yang tiap nilainya disamakan dengan tingkat penerimaan yang berbeda. Ini adalah pendekatan yang digunakan oleh British Council dalam sistem ujian ELTS. Itu mungkin digambarkan sebagai suatu kesan

berdasarkan sistem pemberian tanda. Contoh dari skema nilai yang ditandai ini bisa ditemukan dalam buku karya B.J Carroll (1980b, hal.136).

Pendekatan carroll bagus dalamkonsep seperti memberikan desktipsi yang lebih detail untuk institusi. Masalahnya ia gagal dalam praktek karena tidak melayani para pelajar yang tingkat performa berbeda dipandang dari segi kriteria yang berbeda. Seorang kandidat mungkin diberi nilai 7 untuk ‘kelancaran’, dan nilai 5 untuk ‘keakuratan.

Masalah gagalnya kriteria ini dihindari oleh skema niali yang lebih ‘analitis’, untuk level-level tiap kriteria dan untuk suatu pengukuran yang paling integratif. Metode ini memiliki manfaat lain yang akan menjadikannya lebih mudah untuk enyelesaikan laporan profil dan bisa menunjukkan peran diagnostik dalam menggambarkan kelebihan dan kelemahan hasil tulisan siswa.

Skema nilai analitis tampak seperti suatu alat yang jauh lebih berguna untuk melatih dan menstandarisasi para penguji yang baru. Fancis(1977) menjelaskan bahwa dengan menggunakan skema analitik, menguji isi bisa lebih baik dengan melatih dan menstandarisasi para penguji yang baru tentang kriteria penilaian. Ukuran persetujuan tentang apa yang tiap kriteria harapkan bisa ditentukan, dan para penilai bisa distandarisasi terhadap level yang berbeda dari tiap kriteria ini.

Skema nilai analitis ditemukan dari upaya untuk membuat penilaian lebih objektif. Brooks (1980) menunjukkan bahwa kualitas yang dinilai dengan menggunakan skema nilai analitis di masa lalu sering sulit untuk dipahami. Maka, meskipun skema analitis mungkin memfasilitasi persetujuan diantara para penguji, subjektifitas yang terlibat dalam penilaian pada beberapa skema mungkin sedikit direduksi karena kurangnya ketegasan terhadap kriteria yang dapat dipakai, atau melalui penggunaan yang tidak jelas.

Menentukan Kriteria yang Tepat untuk Menilai Hasil Tertulis:
Test Bahasa Inggrid untuk Tujuan-tujuan Pendidikan (TEEP)

Gagalnya skema analitis di masa lampau telah menjadi pilihan dan gambaran dari krtiteria yang tepat bagi situasi yang diberikan. Dalam model test

TEEP (lihat Weir, 1983a dan Appendikx I) terasa bahwa penilaian contoh-contoh performa tertulis berdasarkan pada kriteria analitis yang tepat yang digolongkan berdasarkan level-level performa yang berbeda.

Data yang melaporkan seleksi kriteria penilaian datang dari survey yang dilakukan terhadap para guru bahasa di sekolah ARELS. Bukti empiris dikumpulkan dari 560 dosen untuk membantu memutuskan kriteria yang bisa digunakan untuk menilai jenis latihan-latihan transfer informasi tertulis yang terjadi dalam konteks akademik.

Sebagai hasil dari investigasi kriteria kerelevanan dan cukup, komposisional, organisasi, kohesi, cukupnya referensi, keakuratan gramatikal, ejaan dan tanda baca terlihat seperti yang paling cocok untuk menilai tugas-tugas menulis.

Untuk menggunakan kriteria yang ‘valid’, satu uapya telah dilakukan untuk membuat skema penilaian analitis yang tiap kriteriannya dibagi menjadi empat level behavioral pada skala 0-3 (lihat tabel di bawah). Level 3 dapat disamakan denga kompetensi minimal. Dalam level ini terasa bahwa siswa mungkin mempunyai sedikit masalah yang berhubungan dengan tugas-tugas menulis. Pada level 2, beberapa masalah yang berhubungan dengan kriteria ini muncul, dan bantuan remedial sebaiknya dilakukan. Level 1 akan mengindikasi bahwa banyak bantuan yang perlu dilakukan terhadap kriteria ini. Level 0 menunjukan hampir semua tidak kompeten dalam respek terhadap pertanyaan.

Skala-skala Penulis Atribut TEEP

A. Relevansi dan Cukupnya Isi

0 - Jawaban hampir tidak mengandung isi yang berhubungan dengan tugas. Total jawabannya tidak cukup.

1- Jawaban mempunyai relevansi yang terbatas terhadap tugas. Memungkinakn adanya celah-celah besar dalam melaporkan topik dan/atau adanya pengulangan yang tidak berarti.

2- Sebagian besar tugas terjawab, meskipun mungkin ada beberapa celah atau informasi yang berlebih-lebihan.

3 - Jawaban sudah relevan dan cukup.

B. Organisasi Komposisional

0 - Tidak adanya organisasi isi yang jelas.

1- Organisasi isi yang ada hanya sedikit.

2- Terdapat beberapa keterampilan organisasional;, tapi belum terkontrol dengan baik.

3- Penggunaan kohesi dan pola internal sudah cukup jelas, keterampila-keterampilan organisasional cukup terkontrol

C. Kohesi

0- Hampir tidak ada kohesi. Tulisan sangat tidak lengkap dan secara virtual tidak mungkin adanya pemahaman terhadap komunikasi yang dimaksud.

1- Kohesi yang tidak memuaskan mungkin menyebabkan sulitnya memahami sebagian besar komunikasi yang dimaksud.

2- Sebagian besar kohesi memuaskan mekipun kadang-kadang kurang memuaskan, mungkin berarti bahwa bagian-bagian tertentu dari komunikasi tidak selalu efektif.

3- Penggunaan kohesi yang memuaskan menghasilkan komunikasi yang efektif.

D. Cukupnya Kosakata

0- kosakata yang dimiliki tidak cukupnya bahkan untuk komunikasi yang paling dasar.

1- Kosakata yang dimiliki untuk tugas tidak cukup. Mungkin leksikalnya sering tidak tepat dan/atau sering ada pengulangan-pengulangan.

2- Ada beberapa kosakata yang tidak cukup untuk tugas. Mungkin ada beberapa leksikal tidak tepat dan/atau adanya pemakaian kata yang berlebih-lebihan.

3- Hampir memiliki cukup kosakata untuk tugas. Hanya kadang-kadang tidak cocok dan/atau berlebihan.

E. Grammar

0- Semua pola gramatikal sering tidak akurat.

1- Grammar sering tidak akurat.

2- Beberapa grammar tidak akurat.
3- Hampir tidak ada grammar yang tidak akurat.

F. Tanda Baca
0- Ketidaktahuan konvensi tanda baca.
1- Standar rendah dari keakuratan tanda baca.
2- Beberapa tanda baca tidak akurat.
3- Hampir tidak ada tanda baca yang tidak akurat.

G. Ejaan
0- Hampir semua ejaan tidak akurat.
1- Standar rendah terhadap keakuratan ejaan.
2- Beberapa ejaan tidak akurat.
3- Hampir tidak ada ejaan yang tidak akurat.

Pertimbangan Berikutnya dalam Membuat Pola Tugas-tugas menulis untuk Dimasukkan ke dalam Rangkaian Test

Nomor Tugas-tugas Menulis

Pembahasan menulis yaitu mengenai bagaimana reliabilitas penilai dapat tercapai. Ada faktor-faktor lain yang memiliki kontribusi pada reliabilitas test. Pertama, beberapa contoh dari karya siswa yang diambil bisa membantu mengontrol perbedaan performa yang mungkin terjadi dari tugas ke tugas.

Baik reliabilitas maupun validitas yang ditingkatkan dengan cara penarikan contoh lebih bnayaka daripada dengan satu komposisi dari tiap kandidat. Finlayson (1951, hal. 132) melihat bahwa “performa dari seorang anak pada satu essay tidak representatif terhadap kemampuannya menulis essay secara umum”. Penelitiaan Vernon dan Milligan (1954, hal.69) juga memperoleh bahwa “ada keraguan yang sangat besar terhadap praktek biasa…mencoba untuk menilai kemampuan bahasa Inggris umum dari satu essay yang dinilai oleh seorang penguji”.

Ebel (1972) menunjukkan bahwa lebih banyak contoh dari tulisan siswa pada suatu test, maka hasilnya akan lebih reliabel. Ebel menguraikan bagaimana skor test terdiri dari dua unsur: skor yang benar dan skor yang salah.

Murphy (1978) juga berpendapat bahwa faktor penting dalam menentukan reliabilitas yang bermacam-macam dari delapan ujian GCE di bawah ini:

> Beberapa nilai untuk individu yang berkontribusi dalam nilai-nilai ujian akhir. Efek peningkatan reliabilitas yang dialkukan dengan cara memiliki lebih banyak bagian dari ujian ditunjukkan oleh kasusu bahasa Inggris level "A". Observasi ini konsisten terhadapa prinsip yang telah ditentukan yang mana kombinasi ukuran-ukuran yang tidak reliabel menjadi lebih reliabel dari ukuran-ukuaran individu itu sendiri.

Jacobs dkk (1981, hal.15) berpendapat bahwa:

> Sebaiknya memperoleh paling tidak dua komposisi dari tiap siswa. Bantuan-bantuan ini memastikan bahwa test tersebut melakukan penarikan contoh representatif dari kemampuan penulis, dengan mereduksi beberapa efek dari variasi performa individu dari topik ke topik atau dari satu periode test yang lain… pengalaman kita yang lainnya mengaharapkan dua tugas menulis yang dirumuskan dengan hati-hati cukup memungkinkan bagi kebanyakan situasi ujian.

Nyatanya lebih banyak contoh tulisan siswa yang diambil lebih baik dari ini akan menjadi tujuan reliabilitas dan validitas, menjadikan tiap contoh memperkirakan kemampuan yang bisa dipertanggungjawabkan.

Pilihan Pertanyaan

Seperti seleksi topik, penting untuk memastikan bahwa siswa mampu menulis sesuatu dengan topik yang diberikan. Apakah berarti membiarkan pemilihan topik itu merupakan suatu keputusan yang penting yang harus dilakukan, untuk itu bisa mempengaruhi reliabilitas test.

Jacobs dkk (1981, hal.1) menyatakan:

> Untuk evaluasi skala besar, sebaiknya semua siswa menulis toik yang sama karena dengan membiarkan memilih topik akan memasukkan begitu banyak perbedaan yang tidak terkontrol kedalam test, artinya apakah perbedaan skor yang diobservasi harus menjadi perbedaan-perbedaan nyata dalam kecakapan menulis atu topik-topik yang berbeda? Tidak ada dasar reliabel bagi perbandingan skor pada suatu test jika semua siswa tidak mengerjakan tugas menulis yang sama; salin itu, konsistensi atau reliabilitas pembaca dalam mengevaluasi test mungkin akan direduksi jika semua bacaan dari sesi penskoran tunggal bukan merupakan topik yang sama.

Heaton (1975) berpendapat bahwa menyediakan pilihan berarti siswa akan menghabiskan banyak waktu untuk mencoba memilih topik dari beberapa alternatif yang diberikan. Dimana test-test yang dilakukan dengan waktu yang terbatas, memaksa siswa untuk menulis denga topik yang sama mungkin juga ber*Manfaat* bagi kandidat yang tidak jelas. Jacobs dkk (1981, hal.17) menyimpulkan:

> Mengingat masalah-masalah yang berhubungan dengan penyediaan pilihan topik, alternatif terbaiknya, jika keterampilan memilih topik bukan diantara tujuan-tujuan test, akan terlihat mengharuskan semua siswa untuk menulis dengan topik yang sama, dan untuk memberi mereka lebih dari satu kesempatan untuk menulis.

Dengan mendasarkan tugas-tugas menulis pada teks tertulis atau lisan yang diberikan kepada kandidat atau stimuli non-verbal, mungkin untuk memastikan bahwa pengetahuan subjek semunya mulai sama paling tidak yang berkenaan dengan informsi yang tersedia untuk mereka. Semuanya dibutuhakn untuk menulis dengan topik yang sam, tapi mereka akanmenulis dengan topik yang berbeda-beda.

Waktu yang Diberikan bagi Tiap Tugas Menulis: Percabangan Limit Waktu

Jacobs dkk (1981, hal.17) menunjukkan perlunya memberi pertimbangan kepada tujuan test menulis:

> Apakah test hasil perkemmbangan langsung dari aktivitas belajar tertentu, mungkin termasuk, revarasi untuk omposisi test (embaca buku-buku tertentu atau melakukan penelitian topik yang ditentukan, mempraktekan topik yang sama tau mode yang sama di dalam kelas), atau apakah test dadakan, yang memusatkan pada hasil gubahan, daripada proses gubahan?

Jacobs dkk (1981, hal.17) menunjukan beberapa cabang dari perbedaan ini:

> Test dadakan dengan waktu terbatas bisa mulai dengan memberi sumber-sumber penulis kepada semua proses menulis, dengan hasil yang mirip dengan apa yang biassa penulis lakukan pada proses menulis. Penting untuk mengingat limitasi waktu.

Waktu yang tepat bagi penyesuaian tugas-tugas menulis orientasi hasil dalam ujian biasa. Jacobs dkk (1981, hal.18) berpendapat bahwa:

> "Test komposisi yang diberikan bersama dengan rangkaian pengukuran lain harus membatasi waktu jika semua waktu test menjadi praktis dan tidak mengenalkan beberapa perbedaan yang pasti akan membosankan bagi peserta ujian… kita memberi batasan waktu 30 menit untuk test komposisi yang diberikan sebagai bagian dari Test Michigan dan waktu yang diberikan itu cukup untuk menghasilkan contoh kemampuan menulis mereka.dalam penelitian mereka (hal.19), mereka menemukan bahwa"Dengan test komposisi selama 30 menit… tapi sebagian besar siswa dengan kemampuan level dasar umumnya bisa menulis sekitar satu halaman atau lebih."

5.3.4 Kesimpulan

Komponen menulis dan beberapa test akan pada tugas-tugas menuli yang dikontrol dimana ciri-ciri audiensi, medium, keadaan, dan tujuan bisa lebih spesifik. Perhatian harus diberikan pada perkembangan kriteria penskoran yang cukup dan tepat serta pada para penguji yang dilatih dan standarisasi terhadap penggunaan ni.

5.4 Ujian Speking (Berbicara)

Ujian kemampuan berbicara memberikan cukup kesempatan untuk menemukan kriteria untuk ujian komunikatif, artinya bahwa: tugas-tugas yang dikembangkan dalam paradigma ini akan mempunyai tujuan, akan menarik, dan mempunyai motifasi, dengan efek wasback positif pada pengajaran yang mendahului test; interaksi akan menjadi ciri kunci; akan ada tingkat intersubjektifitas daintara partisipan; hasilnya akan menjadi tidak bisa diprediksi; konteks realistis akan diberikan; dan pengolahan akan dilakukan. Mungkinlebih banyak dari beberapa keterampilan lain. Ada kemungkinan untuk membnagunnya menjadi test karakteristik dinamis dari komunikasi aktual (lihat bagian 3.1).

Masalah-masalah penialian kemampuan berbicara lebih besar dari penilaian menulis karena interaksinya berlalu dengan cepat dan tidak bisa dicek. Tugas penting bagi pembuat model test harus menentukan aktifitas apa yang harus kandidat tunjukkan, seberapa jauh karakteristis komunikatif dinamis yang

berhubungan dengan aktifitas-aktifitas ini bisa dimasukkan kedalam test, dan dimensi tugas apa yang akan melibatkan kompleksitas, ukuran susunan percakapan referensial dan fungsional untuk diproses dan dihasilkan.

5.4.1 Essay Verbal

Kandidat diminta untuk berbicara selama tiga menit dengan satu topik umum yang ditentukan atau lebih.

Manfaat

1. Kandidat harus berbicara panjang lebar yang memungkinkan untuk menggunakan kriteria-kriteria termasuk kefasihan. Pertanyaan-pertanyaan singkat yang berbeda yang harus siswa jawab sering membatasi susunan kriteria yang dapat diaplikasi.

Kerugian

1. Masalah-masalah yang berhubungan dengan tugas menulis tidak terkontrol bebas menggunakan jenis lisan ini. Topik ditentukan mungkin tidak menarik bagi kandidat dan bukan sesuatu yang meminta kita untuk melakukannya dalam nyata tanpa persiapan.
2. Lebih banyak open-ended topik, maka performa yang ada lebih sukses tergantung pada pengetahuan latar belakang dan pengetahuan kultural dan menggunakan faktor-faktor seperti imaginasi atau kreatifitas. Devinisi respon-respon terhadap apa yang diahrapkan dari isi lebih sulit utnuk mempertahankan reliabilitas dalam penilaian.
3. Pengguaan tape recorder dalam tugas ini mungkin menjadi tekanan bagi para kandidat.

5.4.2 Presentasi Lisan

Kandidat diharapkan untuk berbicara singkat dengan topik yang telah dia siapkan sebelumnya. Berbeda dari “Essay Lisan”.

Manfaat

1. Sangat efektif untuk membuat kandidat menceritakan dirinya sendiri. Dalam test TEEP ini diharapkan sebagai latihan, tapi diketahui bahwa satu menit yang diberikan kepada kandidat untuk berbicara tentang kehidupan pribadinya memberikan semua indikator yang baik dari kecakapan bahasa lisannya yang berkenaan dengan kriteria yang digunakan dalam menilai semua tugas lain. Apa yang penting dalam menilai hasil berbicara memperoleh contoh sufisien dari ucapan kandidat bagi penilaian yang pantas.
2. Mengintegrasi aktifitas dengan mendengarkan atau membaca teks tugas lisan bisa dicocokkan dengan tugas kehidupan nyata yang kandiidatnya harus perform damam situasu target.

Kerugian

1. Jika kandidat mengetahui topik dengan baik sebelumnya, dia bisa mempelajarinya denga baik. Jika waktu yang diberikan untuk persiapan sedikit kemudian dia menghadpi masalah yang akan diuji mungkin pengetahuan bukan sbagai kemampuan linguistik. Jika tugas dihubungkan dengan membaca berdasrkan teori untuk memastikan bahwa semua kandidat memiliki informasi yang biasa kemudian dia diahadapkan dengan masalah membanca yang mungkin mengganggu nilai.
2. Keragaman interpretasi mungkin akan menimbulkan masalah dalam penilaian.

5.4.3 Wawancara Bebas

Jenis wawancara ini yaitu percakapan mengembangkan model yang tidak berstruktur dan tidak ada prosedur-prosedur yang ditentukan sebelumnya.

Manfaat

1. Karena validitas permukaan dan isinya, wawancara merupakan alat untuk menguji keterampilan lisan para kandidat.

2. Wawancara bebas yaitu mirip percakapan yang agak lama dan petunjuk yang diberikan untuk melakukan wawancara. Percakapan mungkin terlihat lebih teliti terhadap pola normal dari interaksi sosial yang tidak formal dalam kehidupan nyata dimana tidak ada agenda yang dirumuskan dengan jelas.

Kerugian

1. Karena tidak ada prosedur-prosedur untuk memperoleh bahasa, performa-performa mungkin berbeda dari satu peristiwa keperistiwa lain karna topik-topik yang berbeda mungkin mulai dibicarakan dan perbedaan-perbedaan ini terjadi dengan wawancara.
2. Prosedur ini memerlukan waktu yang banyak dan sulit untuk dilaksanakan jika ada banyak kandidat.

5.4.4 Wawancara Terkontrol

Dalam prosedur ini terdapat prosedur-prosedur yang ditentukan sebelumnya untuk memperoleh performa. Wawancara FSI mirip dengan model ini (lihat Adams dan Frith, 1979 dan Wilds, 1975).

Manfaat

1. Kemungkinan besar para kandidat diberi pertanyaan yang sama dan oleh karena itu lebih mudah untuk membandingkan performa tiap kandidat.
2. Prosedur ini mempunyai tingkat yang lebih tinggi dari validitas isi dan permukaan daripada teknik-teknik lain selain dari latihan-latihan role play dan celah informasi di UCLES/RSA pada keterampilan komunikatif bahasa Inggris (lihat Appendix III).
3. Dengan latihan dan standarisasi yang cukup dari penguji terhadap prosedur-prosedur dan skala-skala yang digunakan, figur-figur reliabilitas yang dapat dipertanggungjawabkan dapat tercapai dengan mmenggunakan teknik ini. Clark dan Swinton (1979) melaporkan rata-rata reliabilitas intra-rater 0,867 dan reliabilitas inter-rater 0,75 untuk wawancara jenis FSI.

4. Wawancara lisan yang efektif bisa terjadi ketika kandidat diwawancara dan dinilai oleh ahli bahasa dan ahli subjek yang telah distandarisasi.

Kerugian

1. Salah satu kekurangan wawancara yaitu bahwa ia tidak bisa mencakup kandidat-kandidat yang mungkin mendapatkan dirinya sendirinya terlibat di dalamnya bahkan dimana performa-performa level target yang terbatas seperti pada kasus FSI. Dalam wawancara, sulit untuk meniru semua sifat komunikasi kehidupan nyata seperti timbal balik, motivasi, tujuan dan ketetapan.
2. Bahkan ketika prosedur-prosedur untuk memperoleh performa yang ditentukan sebelumnya masih tidak ada jaminan dimana para kandidat akan diberi pertanyaan yang sama dengan cara yang sama bahkan dengan penguji yang sama.

5.4.5 Transfer Informasi: Deskripsi Urutan Gambar

Kandidat ahrus memahamipanel gambar-gambar yang menggambarkan rangkaian peristiwa secara kronologis dan harus menceritakan kisah pada masa lalu berdasarkan gambar. Sebelumnya waktu diberikan kepada kandidat untuk mempelajari gambar-gambar tesebut.

Manfaat

1. Tugas yang diperlakukan para kandidat sudah jelas. Tidak mengahruskan mereka untuk membaca dan mendengarkandan oleh karen itu menghindari kritik kontaminasi penilaian yang memberikan gambar-gambar yang berat sebelah secara kultural atau edukasional.
2. Metode ini akna menajdi prosedur yang efisien dan salah satu dari yang ada untuk membuat kandidat memberikan contoh percakapan yang berhubungan dan membiarkan aplikasi kriteria dalam penilaian. Juga berguna untuk memperoleh kemampuan kandidat untuk menggunakan bentuk-bentuk

gramatikal seperti bentuk past tense untuk kalimat tidak langsung (melaporkan).

3. Karena semua kandidat didesak oleh informasi yang diberikan oleh ambar-gambar yang memberikan perbandingan terhadap kandidat yang tidak dipengaruhi oleh pengetahuan latar belakang dan pengetahuan kultural yang telah diberikan.
4. Nilai teknik tergantung pada gambar-gambar yang jelas dan tidak ambigu dan bebas dari bias kebudayaan dan pendidikan. teknik ini jelas dan disukai oleh dewan pengurus ujian sekolah British. Dalam studi dengan format yang sesuai untuk komponen lisan bagi TOEFL (Clark dan Swinton, 1979) ini terbukti menjadi satu dari format-format yang paling efektif dalam test eksperimental.

Kerugian

1. Keaslian tugas ini terbatas meskipun bisa dikatakan mewakili situasi yang harus menggambarkan sesuatu yang terjadi yaitu rutinitas informasional. Ini mungkin benar-benar menjadi fungsi yang paling penting dalam beberapa kesempatan. Ia memberi tahu sedikit tentang kemampuan kandidat untuk berinteraksi secara lisan.
2. Jika kualitas gambar tidak sempurna maka kandidat tidak mungkin memiliki kesempatan untuk menunjukkan performa terbaiknya. Perbedaan-perbedaan dalam interpretasi mungkin juga menunjukkan ketidak reliabelan terhadap penilaian.

5.4.6 Transfer Informasi Pertanyaan-pertanyaan dari Satu gambar

Penguji memberikan sejumlah pertanyaan tentang isi ganbar yang telah kandidaat pelajari. Pertanyaan-pertanyaan mungkin diharapkan mencakup pikiran-pikiran dan sikap orang-orang di dalam gambar kemudian mendiskusikan perkembangan-perkembangan dari apa yang digambarkan.

Manfaat

1. Mungkin ada *Manfaat* yang dipertimbngkan dalam meneliti teknik ini, yang telah menunjukkan peran yang ada dalam komponen lisan dari test bahas Inggis PLAB bagi para dokter di luar negeri. Dalam PLAB para kandidat diberi petunjuk slide, x-rays, gambar-gambar kondisi medis, dan lain-lain, kemudian diminta untuk memberi pendapat tentang petunjuk itu seperti menjawab pertanyaan-pertanyaan yang berhubungan dengan petunjuk tersebut.

Kerugian

1. Kandidat merupakan pelaku satu-satunya dalam peran responden dan menghilangkan kesempatan untuk memberi pertanyaan. Kriteria tibal-balik, ciri normal dari sebagian besar interaksi lisan tidak diketahui.
2. Gambar-gambar harus jelas dan tegas untuk alasan-alasan yang telah diuraikan pada pembahasan urutan gambar-gambar. Jika sebagian besar kandidat harus diuji selama beberapa hari maka pertanyaan tentang test akan muncul jika gambar-gambar yang sama harus digunakan. Tapi jika gambar-gambar berbeda terpaksa digunkan maka persoalan tentang hal yang bisa diperbandingkan harus dihadapinya.

5.4.7 Tugas-tugas interaksi

Celah Informasi Siswa-siswa

Dalam tugas ini biasanya para siswa bekerja berpasangan dan masing-masing hanya diberi bagian informasi yang penting untuk menyelesaikan tugas. Mereka harus menyelesaikan tugas dengan mencari informasi yang hilang dari yang lain. Para kandidat harus berkomunikasi untuk mengisi celah informasi dalam situasi yang bermakna.

Sertifikat UCLES/RSA untuk keterampilan komunikatif bahasa Inggris memiliki contoh-contoh realistis tertentu dari ini (lihat Appendix III). Seperti perkembangan dari interaksi lawan biscara in muncul setelah diskusikan dan para kandidat harus melaporkan kesimpulan yang diambil dan memberikan alasan keputusan-keputusan yang diambil.

Manfaat

1. Ada beberapa tugas test yang menggambarkan tindakan komunikasi yang lebih baik untuk memenuhi sebagian besar krtiteria yang diberikan oleh Morrow (1979) bagi apa yang menjadikannya komunikatif. Para kandidat akan bebas untuk memilih partne mereka berinteraksi dengan orang-orang yang mereka kenal dan meras senang berkomunikasi dengan mereka.
2. Seperti ciri interaksi yang normal mereka bisa menggunakan bentuk-bentuk pertanyaan, memperoleh informasi, membuat pertanyaan, meminta klarifikasi dan menguraikannys dengan kata-kata sendiri agar sukses dalam test.
3. Tugas itu sangat interaktif dan menjadi lebih pendek dari sebagian besar tugas yang lain yang menggambarkan komunikasi nyata.

Kerugian

1. Ada satu masalah jika seseorang partisipan mendominasi interaksi dari kandidat lainnya pada kesempatan yang terbatas untuk menunjukkan potensi komunikasinya.
2. Sama halnya dengan jika ada perbedaan besar pada kecakapan antara dua orang mungkin mempengaruhi performa dan pertimbangan yang dibuat.
3. Juga ada masalah jika sesorang kandidat lebih tertarik pada suatu topik atau tugas seperti interaksi yang mungkin menjadi suatu hasil.
4. Performa para kandidat dinilai dalam situasi, dan harus memperkirakan kemampuan performa mereka dalam situasi lain.
5. Juga ada ketidakleluasan praktek seperti waktu yang tersedia, kesulitan-kesuliatan administrasi.

Celah Informasi Siswa-Penguji

Untuk menghindari kemungkinan dariketidak seimbangan pada kontribusi kandidat terhadap interaksi, bebrapa dewan pengurus harus mempunyai penguji yang berperan sebagai salh satu partisipan atau menggunakan lawan bicara biasa, seperti gru yang dikenal agar para kandidat merasa nyaman.

Untuk menguji para kandidat secara terpisah mereka bisa diberi diagram, catatan-catatan, dan lain-lain dari informasi yang hilang dan tugas mereka yaitu harus meminta informasi dari penguji.

Manfaat

1. Manfaat utamnya yaitu bahwa ada kesempatan yang lebih besar dimana lawan bicara akan memberi reaksi dengan cara yang sama dengan semua kandidat yang memberikan perbandingan performa meeka yang lebih pantas.

Kerugian

1. Berinteraksi dengan seorang guru sering menjadi tugas yang lebih menakutkan bagi para kandidat daripada berinteraksi dengan teman sebayanya.
2. Ada beberapa bukti bahwa ketika penguji menjadi seorang partisipan dalam berinteraksi, kadang-kadang dia kurang hati-hati meniali performanya sendiri dan performa kandidat (Fisher, 1979).

5.4.8 Role-Play

Beberapa dewan pengurus ujian, contohnya AEB dan UCLES/RSA, termasuk role play dimana kandidat diharapkan untuk memainkan salah satu peran dalam interaksi yang mungkin diharapkan dalam dunia nyata. Interaksi bisa terjadi antara dua siswa atau biasanya penguji berperan sebagai salh satunya.*Kerugian*nya yaitu sulit untuk membuat penil;aiannya pada waktu yang bersamaan ketika dia ambil bagian dalam interaksi. Sepeti dalam latihan celah informasi yang melibatkan guru sebagi lawan bicara dan penguji, bahayanya yaitu bahwa nilai yang diberikan akan mewakili gambaran akhir dari performanya sendiri seperti performa siswa.

Manfaat

1. Teknik ini bisa menjadi valid pada persoalan permukaan dan isi untuk situasi-situasi yang bereda dan pengalaman dewan pengurus ujian memberi kesan

bahwa itu merupakan prsktek dan secara potensial merupakan alat-alat yang sangat valid dan reliabel untuk menilai kemampuan kandidat dalam berpartisipasi secara efektif dalam interaksi lisan.

Kerugian

1. Masalahnya yaitu bahwa kemampuan-kemampuan hitrionik dan beberapa kandidat mungkin memperimbnagkan kemurahan mereka atas biaya yang lebih memusatkan perhatian kepada dirinya sendiri. Juga ada masalah pada semua interaksi lisan yaitu bahwa para kandidat sering menggunakan bahasa untuk melaporkan dan mengatakan apa yang akan mereka katakan daripada berperan secara langsung.
2. Paksaan praktek berjalan dengan baik, khususnya dalam ujian-ujian yang bersekala besar. Jika harus menggunakan *role play* yang berbeda maka harus memberikan perhatian besar untuk memastikan bawa mereka berkedudukan sama dengan perminaat-permintaan para kandidat.