KEBERGANTUNGAN SKOR
UJIAN KARANGAN BAHASA CINA
MURID TAHUN ENAM SEKOLAH JENIS KEBANGSAAN

QUEK WENG KIM

Tesis Yang Dikemukakan Kepada
Fakulti Pendidikan, Universiti Malaya
Sebagai Memenuhi Keperluan Untuk
Ijazah Doktor Falsafah

2010

PENGAKUAN KEASLIAN PENULISAN

Nama: QUEK WENG KIM (No. K.P: 620430-02-5425)

No. Pendaftaran / Matrik: PHA 050038

Nama Ijazah: Ijazah Doktor Falsafah

Tajuk Tesis ("Hasil Kerja ini"):

KEBERGANTUNGAN SKOR UJIAN KARANGAN BAHASA CINA MURID TAHUN ENAM SEKOLAH JENIS KEBANGSAAN

Bidang Penyelidikan: Pengukuran dan Penilaian

Saya dengan sesungguhnya dan sebenarnya mengaku bahawa:

(1) Saya adalah satu-satunya pengarang / penulis Hasil Kerja ini;
(2) Hasil Kerja ini adalah asli;
(3) Apa-apa penggunaan mana-mana hasil kerja yang mengandungi hakcipta telah dilakukan secara urusan yang wajar dan bagi maksud yang dibenarkan dan apa-apa petikan, ekstrak, rujukan atau pengeluaran semula daripada atau kepada mana-mana hasil kerja yang mengandungi hakcipta telah dinyatakan sejelasnya dan secukupnya dan satu pengiktirafan tajuk hasil kerja tersebut dan pengarang / penulisnya telah dilakukan di dalam Hasil Kerja ini;
(4) Saya tidak mempunyai apa-apa pengetahuan sebenar atau patut semunasabahnya tahu bahawa penghasilan Hasil Kerja ini melanggar suatu hakcipta hasil kerja yang lain;
(5) Saya dengan ini menyerah kesemua dan tiap-tiap hak yang terkandung di dalam hakcipta Hasil Kerja ini kepada Universiti Malaya ("UM") yang seterusnya mula dari sekarang adalah tuan punya kepada hakcipta di dalam Hasil Kerja ini dan apa-apa pengeluaran semula atau penggunaan apa jua bentuk atau dengan apa jua cara sekalipun adalah dilarang tanpa terlebih dahulu mendapat kebenaran bertulis daripada UM;
(6) Saya sedar sepenuhnya sekiranya dalam masa penghasilan Hasil Kerja ini saya telah melanggar suatu hakcipta hasil kerja yang lain sama ada denga niat atau sebaliknya, saya boleh dikenakan tindakan undang-undang atau apa-apa tindakan lain sebagaimana yang diputuskan oleh UM.

Tandatangan calon Tarikh

Diperbuat dan sesungguhnya diakui di hadapan,

Tandatangan saksi Tarikh

Nama:
Jawatan:

SINOPSIS

KEBERGANTUNGAN SKOR UJIAN KARANGAN BAHASA CINA MURID TAHUN ENAM SEKOLAH JENIS KEBANGSAAN

Tujuan kajian ini adalah untuk mengkaji perkaitan antara kesan relatif tugasan karangan dan pemeriksa serta impak gabungan bilangan tugasan berdasarkan prosedur pemarkahan yang berlainan terhadap kebergantungan skor karangan daripada perspektif teori G. Seramai 120 orang murid Tahun Enam daripada 10 buah sekolah di sebuah daerah di Perak menyertai kajian ini. Dalam kajian G, reka bentuk separa tersarang $p \times (r: t)$ model rawak digunakan. Pemeriksa berpengalaman seramai 12 orang dibahagikan secara rawak ke dalam empat kumpulan untuk menjalankan pemarkahan dalam dua sesi yang berlainan. Analisis komponen-komponen varian menunjukkan kesan karangan dan pemeriksa adalah bersandar pada prosedur pemarkahan yang berlainan. Dalam kajian D, berdasarkan reka bentuk yang sama, kebergantungan skor bagi prosedur pemarkahan yang berlainan telah dilaporkan. Dapatan kajian menunjukkan pertambahan bilangan tugasan adalah lebih berkesan daripada pertambahan pemeriksa untuk meningkatan kepersisan pengukuran. Dapatan kajian juga menunjukkan aspek penggunaan bahasa dan mekanis (terutamanya dengan kaedah analitik) memerlukan kombinasi bilangan karangan dan pemeriksa yang kurang untuk mencapai pekali G yang tinggi berbanding dengan aspek kandungan dan organisasi manakala kaedah analitik secara relatif mempunyai kelebihan berbanding dengan kaedah holistik. Berdasarkan model gabungan, semua prosedur pemarkahan dapat mencapai pekali G yang melebihi .80 hanya dengan kombinasi tugasan (faset tetap) dan pemeriksa tunggal. Pengkaji mencadangkan

bahawa pentaksiran penulisan pada peringkat sekolah rendah terutamanya ujian karangan Bahasa Cina harus mempertimbangkan penggunaan kaedah analitik dan memberi penekanan ke atas pentaksiran aspek penggunaan bahasa dan mekanis. Selain itu, untuk mencapai tahap kebergantungan yang tinggi berdasarkan kajian ini, sekurang-kurangnya gabungan tiga tugasan karangan dan tiga pemeriksa disarankan.

SYNOPSIS

THE DEPENDABILITY SCORE OF CHINESE ESSAY TEST OF YEAR SIX PUPILS IN NATIONAL TYPE SCHOOL

The purpose of the study is to examine the relationship of the relative effects of writing tasks and raters, and the combined impact of the numbers of tasks and raters based on different rating procedures on the dependability of writing score from the perspective of G-theory. A total of 120 Year Six pupils from 10 schools at one of the district in Perak have participated in the study. In generalizability study (G-study), random model partially nested $p \times (r: t)$ design is employed. Twelve experienced raters are randomly divided into four groups to conduct the marking on two different sessions. Analysis of variance components shows that the effects of tasks and raters are dependent on the different rating procedures. In decision study (D-study), based on the same design as G-study, dependability score for different rating procedures is reported. The findings indicate that the increase of the numbers of writing tasks is more efficient than that of raters in increasing the accuracy of measurement. The results of this study also indicate that aspect of language use and mechanics (especially combined with analytic method), require the combination of less number of tasks and raters to achieve high generalizability coefficient compared to aspect of content and organization while analytic method has the edge on holistic method. Based on the mixed model, all rating procedures are managed to obtain the generalizability coefficient of more than .80 with only a single combination of

writing task (fixed facet) and rater. It is suggested that writing assessment in primary school level especially Chinese essay test should consider in employing analytic method and emphasize on assessing of language use and mechanics. Besides, to achieve high dependability of writing score based on the current study, at least the combination of three writing tasks and three raters are recommended.

PENGHARGAAN

Saya amat bersyukur kepada Tuhan yang telah memberkati saya dan memberi saya kekuatan untuk menempuhi pelbagai cabaran dan dugaan diri sehingga dapat menyempurnakan kajian ini. Jutaan terima kasih yang tidak terhingga saya sampaikan kepada isteri tercinta, Wat Jin dan anak tersayang Long Xiang yang selalu memberi kerjasama, dorongan, sokongan moral dan mendoakan kejayaan saya.

Ribuan terima kasih saya tujukan kepada Bahagian Tajaan Pendidikan yang telah memberikan saya hadiah biasiswa, Bahagian Perancangan dan Penyelidikan Dasar Pendidikan, Kementerian Pelajaran Malaysia yang telah meluluskan permohonan untuk menjalankan kajian ini, Jabatan Pelajaran Perak yang membenarkan kajian ini dijalankan di SJKC di negeri tersebut, guru-guru besar SJKC yang terlibat dalam kajian ini terutamanya En. Ong dan guru-guru SJKC yang sudi melibatkan diri dan meluangkan masa untuk menyediakan maklumat dan data yang dikehendaki. Ribuan terima kasih juga diucapkan kepada rakan-rakan seperjuangan daripada Lembaga Peperiksaan Malaysia atas segala bantuan, sumbangan idea yang kritis dan cadangan yang membina.

Saya ingin mengambil kesempatan ini merakamkan ribuan teriam kasih kepada guru-guru yang memberikan segala kerjasama dan bantuan kepada saya semasa kajian ini dijalankan terutamanya sebagai ahli panel dalam penilaian tugasan karangan, pemurnian instrumen pemarkahan, penterjemahan instrumen pemarkahan dan pemarkahan skrip karangan. Selain itu, saya juga mengucapkan ribuan terima kasih kepada pakar-pakar pentaksiran kerana kesanggupan dan kesudian mereka untuk

menilai instrumen pemarkahan yang digunakan dalam kajian ini. Terima kasih juga disampaikan kepada murid-murid Tahun Enam SJKC yang sudi menjadi sampel kajian saya dan pegawai-pegawai dan kerani-kerani yang bertugas di Fakulti Pendidikan dan Perpustakaan Universiti Malaya yang sentiasa memberi kerjasama sepanjang pengajian saya.

Setinggi-tinggi penghargaan dan terima kasih dirakamkan kepada penyelia yang dihormati, Dr. Shahrir Jamaluddin atas segala bimbingan, nasihat dan pandangan yang membina yang tidak ternilai sepanjang masa penulisan tesis ini. Saya juga amat terhutang budi kepada Profesor Dr. Zulkifli B. Haji Manaf yang memberi galakkan dan dorongan agar saya meneruskan kajian ini.

Akhir sekali, tidak dapat dilupakan juga ingin saya merakamkan setinggi-tinggi penghargaan kepada kawan-kawan, orang perseorangan dan pihak-pihak tertentu yang tidak disebutkan nama satu per satu yang juga turut memberi sumbangan ke arah penyempurnaan tesis ini, mereka akan senantiasa kekal dalam sanubari saya.

Quek Weng Kim
Seri Kembangan,
Selangor Darul Ehsan.

KANDUNGAN

Muka surat

BAB III KAEDAH KAJIAN

BAB IV KEPUTUSAN KAJIAN

BAB V PERBINCANGAN, IMPLIKASI DAN CADANGAN

SENARAI RAJAH

SENARAI JADUAL

SENARAI SINGKATAN

bil	bilangan
bhs	bahasa
DV	variabel bersandar
IV	variabel tak bersandar
JPP	Jabatan Pelajaran Perak
kpd	kepada
KPM	Kementerian Pendidikan Malaysia
PKSR	Penilaian Kemajuan Sekolah Rendah
ULS	*Unweighted least squares*
UPSR	Ujian Pencapaian Sekolah Rendah

SENARAI SIMBOL DALAM TEORI *GENERALIZABILITY*

kajian D	*Decision study* (kajian keputusan)
kajian G	*Generalizability study* (kajian *Generalizability*)
teori G	teori *Generalizability*
x	tersilang dengan
:	tersarang dalam
$\boldsymbol{E}$	jangkaan
p	*person* (selalunya orang atau calon)
r	*rater* (pemeriksa)
t	*task* (tugasan)
$\equiv$	ditafsirkan sebagai
SS (α)	jumlah kuasa dua bagi α
MS (α)	min kuasa dua bagi α
$\boldsymbol{E}MS$ (α)	jangkaan min kuasa dua bagi α
n	saiz sampel bagi sesuatu faset dalam kajian G
n'	saiz sampel bagi sesuatu faset dalam kajian D
N	saiz bagi sesuatu faset dalam semesta cerapan teraku
N'	saiz bagi sesuatu faset dalam semesta generalisasi
Γ	objek pengukuran (selalunya $\Gamma = p$)
$\sigma^2(\alpha)$	komponen varian kesan rawak bagi α dalam kajian G
$\sigma^2(\acute{\alpha})$	komponen varian kesan rawak bagi α dalam kajian D
σ^2 (Γ)	varian skor semesta
σ^2 (δ)	varian ralat relatif
σ^2 (Δ)	varian ralat mutlak
μ_α	populasi dan / atau semesta min skor bagi α
v_α	kesan skor bagi α
$\boldsymbol{E}\rho^2$	pekali G
Φ	indeks kebergantungan atau pekali phi

BAB I

PENGENALAN KAJIAN

1.0 Pengenalan

Bab pendahuluan ini membincangkan perkara-perkara asas yang meliputi latar belakang kajian dan pernyataan masalah yang menjelaskan sebab-sebab kajian ini perlu dijalankan. Di samping itu, tujuan kajian, soalan kajian dan definisi istilah kajian juga dinyatakan. Selain itu, kepentingan dan limitasi yang berkaitan dengan tajuk kajian juga diperjelaskan.

1.1 Latar Belakang Kajian

Dalam bidang pentaksiran atau pengujian bahasa, ujian menulis karangan atau ujian penulisan langsung merupakan salah satu jenis ujian bentuk subjektif yang paling lumrah digunakan. Penggunaan karangan untuk menguji kemahiran menulis dan pencapaian pelajar dalam penguasaan bahasa sudah lama dipraktikkan dalam bidang pendidikan (Hamp-Lyons, 2002; Madaus & O'Dwyer, 1999; Quellmalz, 1990; Read, 1991; Thomas, 2005).

Secara amnya, pentaksiran bahasa membahagikan ujian karangan atau penulisan kepada dua iaitu jenis langsung dan jenis tak langsung (Greatorex & Irenka Suto, 2006). Ujian langsung biasanya mengkehendaki calon menghasilkan sebuah produk karangan atau penulisan asli berdasarkan sesuatu tajuk yang ditetapkan. Skor calon akan ditentukan oleh pemeriksa. Sementara itu, ujian tak langsung atau lebih

dikenali sebagai ujian objektif, sebahagian besarnya adalah terdiri daripada soalan aneka pilihan. Lazimnya, ujian tak langsung menguji calon dari segi penggunaan kosa kata, tatabahasa, kefahaman tentang petikan dan aspek-aspek lain mengenai pengetahuan bahasa dan penggunaan bahasa (McNamara, 2000). Calon biasanya dikehendaki memilih salah satu jawapan yang betul atau paling tepat daripada beberapa pilihan yang diperuntukkan (Madaus & O'Dwyer, 1999).

Pengujian penulisan langsung dianggap mampu mengumpul maklumat yang lebih menyeluruh tentang tahap kemahiran menulis calon. Justeru itu, ia biasanya dianggap antara cara pentaksiran kemahiran menulis yang paling berkesan (Schoonen, 1997). Setidak-tidaknya jika ditinjau dari aspek kesahan muka, bentuk ujian ini jauh lebih berkesan untuk menguji kemahiran menulis pelajar berbanding soalan aneka pilihan yang mempunyai darjah pemiawaian yang tinggi (Schoonen, 2005) kerana bentuk ujian ini lebih mendekati keadaan hidup yang autentik. Oleh itu, tafsiran yang dibuat mengenai tahap kemahiran menulis calon menerusi keadaan ujian seperti ini juga adalah lebih meyakinkan (Bachman & Palmer, 1999), dan memudahkan kita mengaitkan prestasi pencapaian calon masa kini dengan jangkaan prestasinya pada masa depan (Davies et al., 2002).

Selain itu, setelah mengalami beberapa perubahan rangka teori seperti strukturalisme dan formalisme dalam bidang linguistik dan linguistik gunaan pada abad yang lalu, dunia pentaksiran bahasa mulai sedar betapa pentingnya ujian sebagai alat untuk mentaksir kebolehan calon yang sebenar dan menyeluruh. Pentaksiran kemahiran menulis yang dilakukan melalui penulisan karangan dianggap dapat merealisasikan hasrat tersebut. Pada masa kini, perkembangan pentaksiran

bahasa telah melangkah masuk ke zaman lepas moden (lihat Spolsky, 1985 & 1999) dan rata-rata penyelidik dan pakar bahasa berhasrat menerokai pentaksiran langsung yang mempunyai autentisiti yang tinggi (Bachman, 1990). Pendekatan pentaksiran jenis langsung kini memberi penekanan terhadap kesahan dan kebolehpercayaan di samping menguji kemahiran dan pengetahuan bahasa secara serentak.

Hamp-Lyons (1991) telah menyarankan lima ciri yang dimiliki oleh ujian penulisan langsung. Pertama, calon harus menghasilkan sekurang-kurangnya satu teks penulisan berterusan mengandungi paling kurang 100 patah perkataan. Kedua, calon diberi ruang yang agak bebas untuk menghasilkan respons walaupun mereka perlu menjawab mengikut arahan sama ada berdasarkan teks, gambar atau bahan rangsangan yang lain. Ketiga, setiap respons penulisan akan dinilai oleh sekurang-kurangnya seorang atau biasanya dua orang penilai (penilai ketiga diperlukan untuk kes markah yang ekstrim) yang telah mengikuti proses pemiawaian dan menjalani latihan tertentu. Keempat, penilaian dibuat berdasarkan satu panduan piawaian yang mungkin berupa skrip contoh, skala pemarkahan atau huraian yang disediakan tentang jangkaan prestasi pada tahap kecekapan tertentu. Kelima, keputusan tentang penilaian boleh dinyatakan dengan nombor selain menggunakan tulisan dan komen secara lisan atau kombinasi mana-mana cara. Saranan Hamp-Lyons secara am menyentuh tentang sampel karangan, kebebasan respons calon, ketekalan pemeriksa, kriteria pemarkahan dan interpretasi skor dalam ujian penulisan langsung. Namun begitu, aspek kebolehpercayaan skor sampel penulisan juga perlu dipertimbangkan. Menurut Anastasi (1982), kemahiran menulis bagi individu haruslah dinilai melalui beberapa buah karangan berdasarkan tajuk yang berlainan, yang dihasilkan pada hari

yang berlainan dan juga dinilai oleh pemeriksa yang berlainan demi menjamin kebolehpercayaan skor.

Wood (1991, p. 233) memaparkan hubungan ujian penulisan langsung dan ujian penulisan tak langsung berdasarkan konsep kebolehpercayaan dan kesahan. Menurut beliau, ujian tak langsung mempunyai kelebihan dari segi kebolehpercayaan skor tetapi kekurangan bukti yang kukuh dari segi kesahan manakala ujian langsung boleh menyumbang kepada kesahan skor tetapi menjejaskan kebolehpercayaan skor. Pakar pendidik dan pemeriksa juga menganggap bahawa skor ujian penulisan langsung mempunyai kesahan yang lebih tinggi dan impak positifnya dari segi pengajaran melebihi ujian jenis tak langsung (Milanovic, Saville, & Shuhong, 1993). Bagaimanapun, ujian penulisan langsung tidak boleh diandaikan lebih sah daripada ujian tak langsung misalnya soalan aneka pilihan secara tabii (Davies et al., 2002). Selain itu, bilangan soalan karangan yang boleh diuji adalah terhad memandangkan keterbatasan peruntukan masa dalam sesuatu ujian.

Sungguhpun ujian penulisan langsung mempunyai kelebihan, namun ia juga menghadapi masalah-masalah tertentu. Sudah menjadi hakikat bahawa faktor kemanusiaan telah menyisipi segala aspek ujian subjektif. Justeru itu, ujian penulisan langsung sebagai salah satu bentuk ujian subjektif adalah terdedah kepada pelbagai sumber ralat (Hamp-Lyons, 1990) yang boleh mempengaruhi ketepatan skor ujian. Banyak kajian telah menunjukkan bahawa pentaksiran dengan menggunakan sampel karangan cenderung menghasilkan kebolehpercayaan skor yang rendah (Brown, 2007; Chen et. al., 2007; Coffman, 1966; Liu & Zhang, 1998; Schoonen, 2005) kerana sebahagian besar varian skor dalam pentaksiran karangan mengandungi ralat

sistematik yang tidak berkaitan dengan kebolehan menulis yang hendak ditaksir (Coffman, 1971a). Kajian-kajian lepas telah menunjukkan isu kebolehpercayaan skor yang rendah dalam pentaksiran karangan adalah berkaitan dengan kebolehpercayaan skor pemarkahan dan kebolehpercayaan skor tugasan karangan (lihat Lampiran A). Pakar-pakar pentaksiran juga menyedari akibat yang perlu ditanggung daripada penggunaan ujian penulisan karangan kerana dalam pentaksiran dan pengujian karangan, sumber varian yang mungkin timbul adalah berbagai-bagai, misalnya faktor tugasan (karangan berunsur naratif, pendedahan dan sebagainya), faktor pemeriksa, kaedah pemarkahan (contohnya kaedah holistik dan analitik), dan aspek pemarkahan (iaitu pengenalpastian dan pembahagian aspek kemahiran menulis seperti aspek kandungan dan organisasi serta aspek penggunaan bahasa dan mekanis). Selain itu, tajuk karangan (sama ada terdapat pilihan tajuk atau tidak), mod ujian karangan (misalnya mod pensel kertas atau komputer), faktor persekitaran ujian, masa ujian dan sebagainya juga mungkin menghasilkan keberubahan skor (*variability score*) dan mempengaruhi kebolehpercayaan skor. Pada hakikatnya terdapat banyak lagi sumber varian dalam pentaksiran karangan untuk menguji kemahiran menulis, apa yang dinyatakan di sini merupakan sumber varian yang mungkin dan selalu wujud (lihat Cooper, 1984; Huot, 1990; Schoonen, 2005).

Sama ada sesuatu sumber itu dianggap sebagai punca ralat adalah merupakan isu berkaitan dengan teori dan psikologi kognitif (Schoonen, 2005). Penyelidik tertentu mungkin membahaskan bahawa kemahiran menulis karangan berunsur naratif dan pendedahan adalah kemahiran yang berbeza. Namun begitu, dalam konteks pentaksiran (dan penyelidikan) karangan, secara praktiknya, tugasan karangan yang pelbagai boleh dilihat sebagai faset rawak iaitu sebagai sampel

daripada semesta cerapan teraku (*universe of admissible observations*) yang sama kerana semua jenis karangan yang berbeza wujud di bawah satu tajuk besar pentaksiran karangan menurut spesifikasi ujian (Lee & Kantor, 2005) dan boleh digunakan untuk mengeneralisasi kemahiran menulis pelajar. Ini bermakna tugasan yang pelbagai itu boleh saling bertukaran (*exchangeability*) antara satu sama lain. Dengan kata lain, penyelidik menganggap kemahiran menulis sebagai satu kemahiran tunggal yang diperlukan untuk menghasilkan tugasan karangan yang berlainan (Schoonen, 2005). Breland (1983) juga menyatakan bahawa keadaan kepelbagaian tugasan dan bilangannya adalah tak terhingga iaitu mereka bukan sahaja mempunyai topik yang pelbagai malah rangsangan, sasaran pembaca dan tujuan penulisan juga adalah pelbagai.

Teori ujian klasik dapat mengasingkan varian skor cerapan kepada varian benar dan varian ralat rawak. Namun, bahagian varian ralatnya merupakan konstruk yang tidak dapat dileraikan dengan selanjutnya (Shavelson, Webb, & Rowley, 1989). Justeru itu, teori ujian klasik hanya membenarkan anggaran satu jenis atau sumber ralat pengukuran pada setiap masa. Misalnya, kebolehpercayaan uji dan uji semula menganggap keadaan (*occasions*) sebagai sumber ralat manakala kebolehpercayaan ketekalan dalaman menganggap item sebagai sumber ralat. Justeru itu, teori ini tidak mampu menangani keadaan sumber varian ralat pengukuran yang pelbagai. Sementara itu, teori *Generalizability* atau teori G (Brennan, 2001; Cronbach, Gleser, Nanda, & Rajaratnam, 1972; Cronbach, Rajaratnam, & Gleser, 1963; Shavelson & Webb, 1991) yang biasa digunakan dalam mengkaji kebolehpercayaan dan kesahan pentaksiran berasaskan prestasi mampu menangani keadaan pentaksiran yang

kompleks yang sukar diurus oleh pengukuran kebolehpercayaan kaedah tradisi (Linn, Baker, & Dunbar, 1991; Messick, 1995; Miller & Linn, 2000).

Teori G membolehkan penyelidik menganggar secara berasingan pelbagai magnitud sumber varian ralat yang dipanggil sebagai faset (contohnya, faset tugasan, faset pemeriksa dan faset prosedur pemarkahan) dan interaksi faset-faset berkenaan termasuk juga interaksi faset dengan varian benar di kalangan calon secara serentak (Crowley, Thompson, & Worchel, 1994; Shavelson & Webb, 1991). Variasi di kalangan calon adalah varian benar yang ingin diukur. Ia dipanggil sebagai objek pengukuran mengikut istilah teori G.

Teori G merupakan teori statistik mengenai kebergantungan pengukuran. Kebergantungan merujuk kepada kepersisan generalisasi skor suatu ujian daripada skor cerapan kepada min skor calon yang boleh didapati di bawah semua kondisi pengukuran yang mungkin dan juga pengguna ujian sudi menerimanya (Shavelson & Webb, 1991). Selain itu, Teori G menyediakan satu rumusan pekali untuk menggambarkan tahap kebergantungan pengukuran iaitu pekali G dan indeks kebergantungan (pekali phi). Anggaran tahap kebergantungan boleh dijalankan dalam dua peringkat. Peringkat pertama ialah kajian G. Kajian G akan mengenal pasti dan mentafsir seberapa banyak sumber ralat pengukuran yang berpotensi yang boleh didapati daripada semesta cerapan teraku dan seterusnya memberikan anggaran komponen-komponen varian bagi setiap sumber varian ralat (faset) dan interaksi faset yang mendasari prosedur pengukuran. Peringkat kedua iaitu kajian D yang membolehkan penyelidik menggunakan maklumat daripada kajian G untuk membuat keputusan yang boleh meminimumkan kesan varian ralat untuk tujuan tertentu,

misalnya pemadanan optimum bagi bilangan pemeriksa, bilangan tugasan yang berlainan atau kedua-duanya sekali. Biasanya, strategi atau keputusan yang dibuat melibatkan pertimbangan faktor masa, kos, logistik, kebolehpercayaan skor dan sebagainya. Dengan kata lain, kajian G adalah berkaitan dengan pembangunan suatu prosedur pengukuran manakala kajian D mempergunakan prosedur berkenaan (Brennan, 1992; Shavelson & Webb, 1991).

Selain itu, teori G membolehkan penyelidik mentafsir kebergantungan skor ujian menerusi keputusan relatif (pentaksiran rujukan norma) dan keputusan mutlak (pentaksiran rujukan kriteria) atau kedua-duanya sekali bagi individu. Dalam hubungan ini, teori ujian klasik hanya mempertimbangkan sumber ralat pengukuran tunggal bagi keputusan relatif sahaja (VanLeeuwen, 1997). Keputusan mutlak adalah membandingkan pelajar berdasarkan suatu kriteria yang ditetapkan dan menggunakan skor ujian untuk menggolongkan pelajar ke dalam kumpulan tertentu. Misalnya, pelajar ditempatkan dalam kelas pengayaan bahasa sekiranya skor karangan pelajar melepasi skor *cut-off* yang ditetapkan. Manakala Keputusan relatif dibuat untuk membandingkan individu menurut kedudukan dalam kumpulan berdasarkan skor ujian. Misalnya, guru bahasa menggunakan skor daripada pentaksiran karangan untuk menentukan kedudukan pelajar dalam kelas. Dalam kajian ini, oleh kerana tujuan kajian adalah untuk membandingkan kedudukan calon dalam kumpulan iaitu berdasarkan pentaksiran rujukan norma, maka keputusan relatif sahaja akan dibincangkan.

1.2 Pernyataan Masalah

Isu kebolehpercayaan skor yang rendah dalam pentaksiran kemahiran menulis menerusi penulisan karangan atau pentaksiran langsung merupakan isu yang sering mendapat perhatian serius (Breland, 1983; Coffman, 1966 & 1971a). Tujuan utama pentaksiran dan pengujian karangan adalah untuk menilai pencapaian kemahiran menulis calon. Skor yang diberikan kepada produk karangan dijangka dapat mencerminkan kemahiran menulis calon yang sebenar. Namun begitu, faktor tugasan karangan dan faktor pemeriksa sering merupakan sumber ralat utama dalam skor karangan (Bunch & Litterfair, 1988; Cumming, 1990; Huot, 1990a; Kroll, 1998; Pollitt & Hutchinson, 1987; Schoonen, 2005). Selain itu, prosedur pemarkahan juga merupakan faktor penting dalam mempengaruhi kebolehpercayaan skor bagi karangan calon (Carr, 2000; Chi, 2001; Crehen, 1997; Huot, 1990b; Lehmann, 1993; Swartz et al., 1999). Dalam pentaksiran karangan, faktor-faktor ini sama ada bertindak secara bebas atau berinteraksi antara satu sama lain telah mewujudkan kepelbagaian sumber varian yang sukar ditangani dan seterusnya menyukarkan pentafsiran skor.

Masalah pekali kebolehpercayaan yang rendah merupakan masalah yang sering dihadapi dalam pentaksiran penulisan karangan pelajar (Lee & Kantor, 2005; Liu & Zhang, 1998; Schoonen, 2005). Hasil dapatan kajian lepas telah menunjukkan *generalizability* skor yang rendah merentas tugasan atau jenis tugasan yang digunakan untuk mentaksir kemahiran menulis pelajar (Baker, Abedi, Linn, & Niemi, 1996; Boodoo & Garlinghouse, 1983; Breland, Bridgeman, & Fowles, 1999; Breland, Camp, Jones, Morris, & Rock, 1987; Brennan, Gao, & Colton, 1998; Brown, Hilgers, & Marsella, 1991; Cantor dan Hoover, 1986; Chen, Niemi, Wang,

Wang, & Mirocha, 2007; Cumming, Kantor, Powers, Santos, & Taylor, 2000; Gabrielson, Gordon, & Engelhard, 1995; Lamb, 1987; Lee & Kantor, 2005; Lehmann, 1990; Liu & Zhang, 1998; Powers & Fowles, 1999; Moss, Cole, & Khampalikit, 1982; Schoonen, 2005; Stevens & Clauser, 1996). Ini bermakna skor untuk tugasan penulisan yang berlainan biasanya memaparkan korelasi yang kurang memuaskan. Lehmann (1990) melaporkan korelasi antara .10 hingga .46 untuk sembilan tugasan penulisan yang berlainan dalam kajian IEA mengenai pencapaian penulisan karangan pelajar antarabangsa. Kajian Cantor dan Hoover (1986) ke atas lima bentuk karangan yang berbeza iaitu naratif, pendedahan, deskriptif, laporan informatif dan pemujukan merentas pelajar Gred 3 hingga Gred 8 mendapati bahawa kebolehpercayaan skor antara tugasan berkenaan hanya berada dalam lingkungan .30 hingga .40. Sementara itu, kajian Schoonen (2005) ke atas empat jenis tugasan iaitu meliputi penulisan berunsur deskriptif, penerangan dan pemujukan terhadap pelajar Gred 6 berdasarkan aspek dan kaedah pemarkahan yang berlainan juga menunjukkan pekali G yang agak rendah: aspek kandungan dan penyusunan (kaedah holistik .32; kaedah analitik .21) serta aspek penggunaan bahasa (kaedah holistik .40; kaedah analitik .30) berdasarkan cerapan tunggal. Manakala kajian Liu & Zhang (1998) ke atas karangan bentuk hujahan, naratif dan gabungan bentuk hujahan dan naratif berdasarkan aspek kandungan, organisasi dan bahasa dengan pemarkahan holistik memperlihatkan pekali G yang sangat rendah (.12) berdasarkan cerapan tunggal.

Penyelidik dan pakar teori pernah memaparkan banyak aspek tugasan karangan yang boleh mempengaruhi prestasi penulisan pelajar di samping kemahiran menulis yang sebenar. Aspek-aspek yang disentuh dalam kajian lepas ialah bentuk karangan (contohnya, Quellmalz et al., 1982), latar belakang pengetahuan yang

diperlukan oleh tugasan (contohnya, Ruth & Murphy, 1984), tajuk tugasan (contohnya, Hoetker; 1982; Hoetker & Brossell, 1989; Peterson, 2000), spesifikasi retorik (contohnya, Brossell, 1983; Smith et al., 1985), kesukaran tugasan (contohnya, Hamp-Lyons & Prochnow, 1994; Breland, Lee, & Najarian, & Muraki, 2004) dan sebagainya. Semua kesan tugasan tersebut adalah berpotensi mempengaruhi kebolehpercayaan skor penulisan pelajar. Mereka boleh dianggap sebagai faset atau sumber varian ralat bagi tugasan karangan.

Selain tugasan karangan, keberubahan skor berhubung dengan pemarkahan pemeriksa juga merupakan faktor kritikal yang sering dikaji dalam pentaksiran penulisan. Faktor pemeriksa selalu dianggap sebagai sumber varian utama dalam pentaksiran karangan selain tugasan karangan (Bachman, Lynch, & Mason,1995; Huot, 1990; Kroll, 1998; Schoonen, 2005). Dalam mengkaji pengaruh faktor pemeriksa terhadap kebolehpercayaan skor untuk soalan subjektif (termasuk soalan karangan) dan objektif, dapatan kajian Wang (2001) menunjukkan bahawa faktor pemeriksa merupakan sumber varian utama dalam soalan subjektif terutamanya dalam pemarkahan karangan.

Pemeriksa sering memberi tafsiran yang berlainan terhadap kriteria pemarkahan yang sama dan keadaan ini adalah sukar diselaraskan antara mereka. Kecenderungan pemeriksa melakukan penilaian yang berbeza terhadap mutu karangan selalu dikaitkan dengan perbezaan dari segi latar belakang pengetahuan (Jenning, Fox, Graves, & Shohamy, 1999), latar belakang budaya (Kobayashi & Rinnert, 1996; Shi, 2001 & 2003), kepakaran pemeriksa (Barnwell, 1989; Cumming, 1990; Erdosy, 2004; Schoonen, Vergeer, & Eiting, 1997; Shohamy, Gordon, &

Kraemer, 1992), ketegasan pemeriksa (Lunz, Wright, & Linacre, 1990), latihan pemeriksa (Lumley & McNamara, 1995), proses kognitif pemeriksa (Wolfe, Kao, & Ranney, 1998), sikap prasangka pemeriksa tentang calon dan kriteria pemarkahan tertentu (Kondo-Brown, 2002) dan sebagainya. Di samping itu, pemeriksa mungkin juga memberi penekanan yang berbeza terhadap kriteria pemarkahan dan ini merupakan masalah umum yang dihadapi. Pemeriksa mungkin mendenda kesalahan tatabahasa dan aspek mekanis dengan lebih berat daripada aspek ciri linguistik, kandungan dan organisasi (Bock, 1998). Selain itu, tahap emosi dan kesihatan pemeriksa juga boleh menggugat kestabilan pemberian skor.

Walaupun latihan pemeriksa boleh mengurangkan perbezaan pemarkahan yang ekstrim, namun perbezaan pemarkahan yang signifikan antara pemeriksa masih berlaku (Lumley & McNamara, 1995). Latihan pemeriksa tidak dapat menghasilkan kesan yang sepenuhnya kerana pemeriksa mempunyai persepsi unik dan merasakan bahawa "*they have unique standards, and it is hard for them to alter their standards*" (Lunz & Stahl, 1990, p. 428). Perubahan tingkah laku pemeriksa pada masa atau keadaan yang berlainan juga boleh menjejaskan pemberian skor kepada calon (Congdon & McQueen, 2000).

Kesan tugasan dan kesan pemeriksa besar kemungkinan bersandar kepada kriteria yang dinilai dan cara ia dinilai. Dengan kata lain, aspek pemarkahan dan kaedah pemarkahan mungkin mempengaruhi kesan tugasan dan kesan pemeriksa (Schoonen, 2005). Dalam kajian ini, kaedah pemarkahan yang digunakan untuk menilai sampel karangan adalah kaedah holistik dan analitik manakala aspek pemarkahan yang digunakan adalah aspek kandungan dan organisasi serta aspek

penggunaan bahasa dan mekanis. Kedua-dua aspek berkenaan adalah saling melengkapi.

Andaian bahawa kesan tugasan mungkin bersandar kepada aspek pemarkahan barangkali lebih mudah dijangkakan daripada kesan pemeriksa. Secara rasional, aspek kandungan dan organisasi dalam kemahiran menulis mungkin mempunyai perkaitan yang lebih kuat dengan pengetahuan yang dikehendaki dalam tajuk karangan. Oleh itu, prestasi calon mungkin lebih bergantung kepada ciri-ciri spesifik tugasan karangan. Manakala aspek penggunaan bahasa dan mekanis adalah menjurus kepada penguasaan bahasa yang betul dan berkesan serta ciri-ciri luaran bahasa yang dituntut dalam setiap tugasan, tanpa melibatkan pengetahuan mengenai tajuk karangan. Justeru itu, aspek penggunaan bahasa dan mekanis boleh dianggap kurang berkaitan dengan isi tugasan karangan berbanding dengan aspek kandungan dan organisasi.

Bagimanapun, setakat mana kesan pemeriksa bersandar kepada aspek pemarkahan adalah sukar dijangkakan. Namun begitu, kajian Cumming (1990) telah melaporkan pemarkahan pemeriksa berpengalaman dan tidak berpengalaman bagi aspek kandungan dan aspek retorik organisasi adalah berbeza secara signifikan dan aspek penggunaan bahasa menunjukkan tidak terdapat perbezaan. Schoonen et al. (1997) pula melaporkan pemarkahan pemeriksa tidak berpengalaman kurang boleh dipercayai dari segi aspek penggunaan bahasa berbanding dengan aspek kandungan. Kajian Sultana (2001) pula menunjukkan bahawa semakin lama suatu rubrik pemarkahan karangan yang sama digunakan, semakin kurang ketat pemeriksa dalam pemberian skor. Kajian Eckes (2008) pula memaparkan bahawa pemeriksa

berpengalaman memperlihat kepentingan yang berlainan terhadap kriteria pemarkahan tertentu dan mereka juga boleh digolongkan mengikut kriteria pemarkahan berkenaan. Masalah ini bertambah runcing lagi kerana adalah mustahil membentuk satu rubrik pemarkahan yang dapat merangkumi semua respons yang mungkin dihasilkan oleh calon dalam pentaksiran respons terbuka seperti ujian menulis karangan (Yin & Shavelson, 2004).

Kesan tugasan dan kesan pemeriksa mungkin juga bersandar kepada kaedah pemarkahan yang digunakan dalam pentaksiran karangan. Pemeriksa mungkin memberi tafsiran yang berlainan terhadap tugasan karangan berdasarkan kaedah pemarkahan yang berlainan. Selain itu, perasaan prasangka pemeriksa dalam penilaian karangan mungkin dipengaruhi oleh kaedah pemarkahan yang berlainan iaitu kaedah holistik dan analitik (Hoover & Politzer, 1981). Malah kajian tentang menggunakan skema pemarkahan alternatif untuk menilai kecekapan menulis telah menunjukkan bahawa cara-cara pemarkahan yang nampaknya serasi pada asalnya mungkin menghasilkan keputusan yang berbeza walaupun penilaian dilakukan ke atas set karangan yang sama (Quellmalz, 1990).

Sebenarnya, keberkesanan dan kesesuaian sesuatu kaedah pemarkahan dalam menilai mutu karangan pelajar dan kesannya terhadap kebolehpercayaan skor telah lama mendapat perhatian (Cooper, 1984; Breland, 1983; Spandel & Stiggins, 1981). Pada umumnya, kajian lepas melaporkan bahawa anggaran pekali kebolehpercayaan kaedah pemarkahan analitik adalah lebih tinggi daripada holistik (misalnya, Huot, 1990a; Klein et al., 1998). Jika dilihat dari segi teori, ujian yang mempunyai bilangan item dan tugasan yang lebih banyak biasanya mempunyai kebolehpercayaan yang

lebih tinggi (Allen & Yen, 1979; Linn & Gronlund, 2005). Dalam hubungan ini, pemarkahan analitik dengan bilangan subskala yang lebih banyak mempunyai kelebihan dari segi ketekalan pemarkahan keseluruhan (lihat Brown & Bailey, 1984; Hamp-Lyons, 1991). Namun begitu, Hamp-Lyons dan Kroll (1997) juga mengakui pemarkahan holistik dapat menyamai kelebihan pemarkahan analitik kerana pembinaan subskala yang berfungsi dan dapat membezakan antara satu sama lain dalam pemarkahan analitik adalah kerja yang tidak mudah dilakukan. Ini bermakna subskala yang banyak mungkin menambahkan ralat skor pemarkahan.

Terdapat juga ramai pakar pendidik dan pengukuran menganggap kaedah holistik adalah lebih langsung dan lebih sesuai digunakan untuk mengukur kemahiran menulis pelajar daripada kaedah analitik (Cooper, 1977; Lloyd-Jones, 1977). Selain itu, mereka juga berpendapat bahawa bahasa haruslah dilihat sebagai satu entiti dan tidak sesuai ditaksirkan secara berasingan berdasarkan aspek-aspek tertentu (Arshad Abd. Samad, 2004). Namun begitu, para penyelidik juga mendakwa pemarkahan holistik lebih tertumpu kepada aspek kelebihan daripada aspek kelemahan penulis, sedangkan aspek kelemahan adalah lebih penting dalam membuat keputusan penempatan (Cumming, 1990; Hamp-Lyons, 1990b; Reid, 1993; Cohen, Manion, & Morrison, 2000; White, 1994; Elbow, 1999). Lagipun, pencapaian dan penguasaan aspek-aspek kemahiran menulis adalah berbeza-beza bagi pelajar yang berlainan. Oleh itu, penilaian rujukan kriteria atau maklum balas yang lebih spesifik adalah perlu untuk mengesan kemajuan dan tahap pencapaian pelajar (Bacha, 2001). Ini menunjukkan bahawa pendekatan pemarkahan yang berlainan tidak akan memberi kesahan skor yang sama (Hudson & Veal, 1981). Moss et al. (1982) mendapati julat korelasi antara tugasan dengan pemarkahan global (holistik)

adalah antara .41 hingga .50 manakala bagi pemarkahan analitik adalah antara .30 hingga .73. Ini menunjukkan bahawa kaedah pemarkahan yang digunakan boleh mempengaruhi kebolehpercayaan skor karangan (Breland, 1983; Coffman, 1971a & 1971b). Justeru itu, kaedah pemarkahan adalah relevan dipertimbang sebagai sumber untuk menilai kebolehpercayaan skor merentas tugasan dan pemeriksa.

Kemahiran menulis merupakan kemahiran penting yang perlu dikuasai dalam mana-mana sistem pendidikan kerana kegagalan dalam kemahiran menulis akan menjejaskan keseluruhan mutu akademik (Abdul Shukor Shaari, 2001; Hamp-Lyons & Kroll, 1997; Nadzri Isa, 2003). Malah kemahiran menulis juga dianggap oleh masyarakat antarabangsa sebagai petunjuk tahap literasi sesebuah negara selain kemahiran membaca (Hasuria Omar, 1999). Oleh sebab itu, murid-murid sekolah rendah di negara ini dikehendaki mempelajari pelbagai jenis tugasan karangan berdasarkan pendekatan yang berbeza iaitu merangkumi penceritaan, pendedahan, penerangan, pemujukan, penghujahan dan sebagainya. Begitu juga dalam konteks Sekolah Jenis Kebangsaan Cina [SJK(C)]. Murid-murid diajar agar menguasai pelbagai jenis tugasan karangan setelah memperoleh asas pengetahuan bahasa. Antaranya, karangan jenis pendedahan dan penceritaan atau naratif adalah paling diutamakan kerana kedua-duanya merupakan asas kemahiran menulis yang perlu dikuasai oleh semua murid pada peringkat sekolah rendah. Ini adalah selaras dengan proses pengajaran dan pembelajaran kemahiran menulis yang dinyatakan dalam sukatan dan huraian sukatan pelajaran sekolah SJK(C) (KPM, 1998, 1999, 2003b & 2003c). Dalam pengajaran menulis karangan Bahasa Cina, guru bahasa lazimnya mengajar berdasarkan tajuk karangan yang diberikan dan murid diminta menyiapkan

tugasan secara individu seperti yang biasa dipraktikkan oleh pengajaran kemahiran menulis dalam bahasa ibunda yang lain (Yahya Othman, 2005).

Dalam konteks pentaksiran karangan Bahasa Cina untuk murid Tahun Enam, tahap pencapaian kemahiran menulis mereka bukan sahaja dinilai dalam pentaksiran berasaskan sekolah, daerah dan negeri tetapi juga dalam pentaksiran pusat iaitu Ujian Pencapaian Sekolah Rendah (UPSR). Dalam UPSR, calon dikehendaki menulis dua tugasan karangan yang berlainan tajuk dalam masa satu jam (lihat LPM, 2005). Skrip jawapan calon akan dinilai oleh pasukan pemeriksa yang terlatih dan berpengalaman. Mereka terdiri daripada guru yang mengajar mata pelajaran Bahasa Cina. Untuk tujuan pentaksiran, tajuk karangan bagi jenis pendedahan dan penceritaan biasanya diutamakan kerana kedua-dua jenis penulisan tersebut merupakan asas kemahiran menulis yang perlu dikuasai. Malah dalam ujian menulis karangan pada peringkat sekolah, daerah dan negeri, kedua-dua jenis karangan tersebut juga dijadikan bahan ujian kerana sikap mengajar untuk tujuan peperiksaan (*teaching to the test*) masih tebal dalam pendekatan pengajaran di SJK(C) (Quek, 2003). Oleh sebab keghairahan untuk mengejar kecemerlangan keputusan peperiksan, para pendidik akan mengamalkan apa yang dipraktik oleh badan peperiksan (Khodori Ahmad & Jamil Adimin, 2003). Apakah kebolehpercayaan skor sekiranya hanya dua tugasan karangan digunakan?

Sementara itu, mulai tahun 2005, prosedur pemarkahan untuk ujian menulis karangan Bahasa Cina UPSR telah bertukar kepada prosedur pemarkahan holistik daripada prosedur pemarkahan analitik yang digunakan untuk sekian lama ekoran pertukaran format peperiksaan (LPM, 2005). Berikutan perubahan tersebut, adakah

skor bagi tugasan karangan yang berlainan dan pemarkahan pemeriksa bergantung kepada prosedur pemarkahan yang berlainan? Apakah kombinasi bilangan tugasan dan pemeriksa terhadap kebolehpercayaan skor berdasarkan prosedur pemarkahan yang berlainan?

Faktor tugasan karangan, pemeriksa dan prosedur pemarkahan yang mempengaruhi ketepatan skor calon serta mengancam kebolehpercayaan dan keadilan peperiksaan telah menjadi isu yang selalu diberi perhatian serius sama ada oleh penyelidik pendidikan dalam negara (Abdul Aziz Abdul Talib, 1993; Arshad Abd. Samad, 2004; Azman Wan Chik, 1994; Ng, 1991) mahupun luar negara (Chen et al., 2007; Coffman, 1971b; Lumley & McNamara, 1995). Dalam mana-mana sistem pendidikan yang menggunakan produk karangan untuk mentaksir kemahiran menulis sering digugat oleh masalah kebolehpercayaan skor terutamanya yang berpunca daripada tugasan karangan, pemeriksa dan juga prosedur pemarkahan. Oleh sebab terdapat pelbagai kekangan yang dihadapi dalam penyelidikan dan pentaksiran karangan, biasanya kebanyakan penyelidik hanya memilih satu daripada faktor-faktor tersebut untuk menjalani analisis. Kajian yang menganalisis kombinasi faktor-faktor jenis tugasan karangan, pemeriksa dan prosedur pemarkahan secara serentak masih tidak pernah dijalankan di negara ini.

Dalam kajian ini, pengkaji akan menjalankan kajian empirikal ke atas beberapa sumber varian utama yang mampu mempengaruhi kebergantungan skor karangan calon iaitu tugasan karangan berlainan, pemeriksa, dan prosedur pemarkahan dengan serentak. Pengkaji juga mengandaikan bahawa kesan tugasan karangan yang berlainan dan kesan pemeriksa mungkin bersandar kepada kaedah

pemarkahan (kaedah holistik dan kaedah analitik) dan aspek pemarkahan (aspek kandungan dan organisasi serta aspek penggunaan bahasa dan mekanis) yang berlainan. Dengan kata lain, kesan pemeriksa dan kesan tugasan karangan berlainan mungkin mempengaruhi keberubahan skor, pada masa yang sama kesan kedua-dua faktor ini mungkin juga berubah berdasarkan kaedah pemarkahan dan aspek pemarkahan yang berbeza, dan seterusnya memberi pengaruh yang berlainan terhadap kebergantungan skor karangan. Untuk meninjau keadaan tersebut, kesan sumber varian perlu dianalisis secara kuantitatif dan dilakukan dengan serentak. Dengan itu, keputusan yang diperoleh dapat memberi pemahaman yang lebih mendalam tentang hubungan rumit yang ditunjukkan oleh faktor-faktor yang menghasilkan skor ralat dalam proses pentaksiran mutu karangan.

Untuk menganalisis dan menganggar secara serentak kesan faktor-faktor ralat yang dipilih dengan tepat, pengkaji menggunakan teori G untuk menjalankan kajian (Saville, 2003, p.71). Ini kerana kesesuaian dan kelebihan teori G dalam aspek kajian mengenai ralat pemarkahan dan kebolehpercayaan skor. Analisis teori G (Brennan, 2001; Cronbach et al., 1963; Cronbach et al., 1972; Shavelson & Webb, 1991) yang digunakan meliputi proses kajian *Generalizability* (kajian G) dan kajian Keputusan (kajian D). Kajian G bertujuan mengungkaikan komponen-komponen varian yang terlibat dan menganalisis sumber varian bagi kesan tugasan karangan berlainan, kesan pemeriksa berdasarkan kesan prosedur pemarkahan yang berlainan. Manakala kajian D menggunakan anggaran varian dalam kajian G untuk meninjau gabungan bilangan tugasan karangan dan bilangan pemeriksa berdasarkan prosedur pemarkahan yang berlainan dari segi kebergantungan skor iaitu dari segi perubahan pekali *Generalizability* (pekali G).

1.3 Objektif Kajian

Objektif umum kajian ini adalah untuk mengkaji pengaruh kesan karangan (berlainan tugasan), kesan pemeriksa, dan kesan prosedur pemarkahan yang berlainan terhadap kebergantungan skor karangan dari segi pentaksiran menulis karangan bagi murid Tahun Enam di SJK(C). Karangan berlainan tugasan merujuk kepada karangan berunsur naratif dan karangan berunsur pendedahan. Manakala prosedur pemarkahan yang berlainan merujuk kepada kaedah dan aspek pemarkahan. Kaedah pemarkahan merangkumi kaedah holistik dan analitik. Aspek pemarkahan meliputi aspek kandungan dan organisasi serta aspek penggunaan bahasa dan mekanis. Kajian ini juga meninjau pengaruh kesan kaedah dan aspek pemarkahan yang berlainan terhadap kesan karangan berlainan tugasan dan kesan pemeriksa. Selain itu, impak gabungan bilangan tugasan dan bilangan pemeriksa berdasarkan prosedur pemarkahan yang berlainan juga dikaji.

Kajian ini menggunakan analisis teori G yang merangkumi dua peringkat iaitu kajian G dan kajian D. Kajian G digunakan untuk menganalisis kesan pemeriksa, kesan karangan berlainan tugasan, kesan prosedur pemarkahan yang berlainan menerusi anggaran komponen varian. Manakala kajian D dijalankan untuk meninjau impak gabungan bilangan pemeriksa dan bilangan karangan berdasarkan prosedur pemarkahan yang berlainan. Dalam kajian G, reka bentuk eksperimen dalam teori G iaitu reka bentuk separa tersarang $p \times (r: t)$ model rawak telah digunakan manakala reka bentuk yang sama berdasarkan model rawak dan model gabungan (dengan tugasan ditetapkan) telah digunakan dalam kajian D. Oleh sebab tujuan pentaksiran karangan dalam kajian ini bertujuan untuk membandingkan kedudukan relatif calon iaitu bercirikan interpretasi rujukan norma, maka pekali G

adalah sesuai dilaporkan. Kumpulan sasaran kajian adalah murid Tahun Enam SJK(C) di salah sebuah dearah di Perak. Secara khususnya, objektif kajian ini adalah untuk:

1. menentukan sama ada faktor karangan berlainan tugasan, pemeriksa, aspek pemarkahan dan kaedah pemarkahan mempengaruhi keberubahan skor.
2. menganalisis perkaitan antara kesan karangan berlainan tugasan dan kesan pemeriksa terhadap prosedur pemarkahan yang berlainan dari segi kebergantungan skor berdasarkan kerangka kajian G.
3. menganalisis impak gabungan bilangan tugasan dan bilangan pemeriksa bagi prosedur pemarkahan yang berlainan terhadap kebergantungan skor berdasarkan kerangka kajian D.

1.4 Soalan Kajian

Kajian ini akan menjawab soalan-soalan seperti berikut berdasarkan objektif khusus kajian dalam perkara 1, 2 dan 3:

1 Adakah kemungkinan faktor karangan berlainan tugasan, pemeriksa, aspek pemarkahan dan kaedah pemarkahan mempengaruhi keberubahan skor?

2 Sejauh manakah perkaitan antara kesan karangan berlainan tugasan dan kesan pemeriksa terhadap prosedur pemarkahan yang berlainan dari segi kebergantungan skor karangan berdasarkan kerangka kajian G?

3 (a) Berdasarkan kerangka kajian D, apakah impak bagi bilangan tugasan karangan, bilangan pemeriksa dan gabungan kedua-duanya terhadap kebergantungan skor bagi prosedur pemarkahan seperti berikut:

(i) kaedah holistik serta aspek kandungan dan organisasi

(ii) kaedah analitik serta aspek kandungan dan organisasi

(iii) kaedah holistik serta aspek penggunaan bahasa dan mekanis

(iv) kaedah analitik serta aspek penggunaan bahasa dan mekanis

3 (b) Sejauh manakah impak gabungan bilangan tugasan karangan dan bilangan pemeriksa bagi prosedur pemarkahan yang berlainan terhadap kebergantungan skor berdasarkan kerangka kajian D?

3 (c) Sejauh manakah impak gabungan bilangan tugasan karangan dan bilangan pemeriksa dengan faset tugasan ditetapkan bagi prosedur pemarkahan yang berlainan terhadap kebergantungan skor berdasarkan kerangka kajian D?

1.5 Definisi Istilah

Dalam konteks kajian ini, istilah-istilah yang digunakan mendukung makna seperti berikut:

1.5.1 Kebergantungan

Kebergantungan suatu skor ujian boleh ditafsirkan sebagai kepersisan generalisasi skor suatu ujian daripada skor cerapan kepada min skor seseorang calon yang diperolehnya dalam semua kondisi pengukuran yang mungkin terdapat dan juga pembuat keputusan atau pengguna ujian sanggup menerimanya (lihat Shavelson & Webb, 1991, p.1).

1.5.2 Faset pengukuran (*Facet of measurement*)

Faset pengukuran merujuk kepada sumber ralat pengukuran dalam teori G (Suen, 1990). Faset merupakan satu himpunan kondisi pengukuran yang serupa (Brennan, 1992, p. 2). Kes-kes berasingan dalam suatu faset merupakan paras faset berkenaan (Suen, 1990, p. 47). Misalnya dalam ujian karangan, dua tugasan

dikemukakan dan respons calon dinilai oleh tiga orang pemeriksa, maka faset tugasan dan pemeriksa adalah dua dan tiga paras masing-masing. Faset boleh jadi rawak atau tetap bergantung kepada kehendak pembuat keputusan dalam membuat generalisasi kepada populasi yang lebih besar yang serupa dengan populasi asal.

1.5.3 Faset rawak (*Random facet*)

Sesuatu faset dianggap rawak sekiranya kondisi-kondisi daripada faset berkenaan yang digunakan dalam pengukuran adalah sampel rawak daripada semesta yang lebih besar di mana pembuat keputusan ingin membuat generalisasi (Brennan, 1992; Shavelson & Webb, 1991). Misalnya, dalam kajian ini pemeriksa adalah dianggap sebagai sampel rawak daripada semesta pemeriksa teraku yang lebih besar.

1.5.4 Faset tetap (*Fixed facet*)

Sesuatu faset dianggap tetap sekiranya kondisi faset yang diandaikan menghabiskan semua kondisi yang mungkin terdapat dalam semesta berkenaan (Shavelson & Webb, 1991, p. 65). Oleh itu, tiada generalisasi yang boleh dibuat ke atas feset berkenaan. Suatu faset juga dianggap sebagai tetap apabila kondisi faset tersebut wujud dalam bilangan yang terhad, dan pembuat keputusan hanya berminat untuk mengkaji prestasi bagi beberapa kondisi atau kondisi tertentu atau semua kondisi yang dicerap (Shavelson & Webb, 1991, p. 129). Faset tetap juga dianggap sebagai faset pemiawaian dan merupakan sebahagian daripada objek pengukuran (Shavelson & Webb, 1991, p. 47)

1.5.5 Kajian *Generalizability* atau kajian G (*Generalizability study*)

Kajian G dalam konteks Teori G adalah satu kajian di mana data dikumpul dan dianalisis di bawah lingkungan semesta cerapan tertentu yang boleh diterima berdasarkan pilihan penyelidik. Tujuan kajian G adalah untuk meleraikan dan menganggar varian-varian berkaitan dengan faset pengukuran. Berdasarkan komponen varian yang didapati, penyelidik akan membentuk prosedur pengukuran yang memanfaatkan dan berkesan (Brennan, 1992, p. 3).

1.5.6 Karangan Bahasa Cina

Karangan Bahasa Cina yang dimaksudkan dalam kajian ini hanya merujuk kepada karangan bentuk naratif dan bentuk pendedahan.

1.5.7 Kebolehpercayaan skor

Kebolehpercayaan skor yang dimaksudkan dalam kajian ini adalah berkaitan dengan emaun varian ralat dan kebolehan generalisasi yang dihasilkan oleh skor karangan bentuk naratif dan bentuk pendedahan. Ia dianggar melalui kesan karangan berbeza bentuk, kesan pemeriksa dan kesan prosedur pemarkahan yang berlainan dengan menggunakan teori G.

1.5.8 Keputusan mutlak (*Absolute decisions*)

Ia merupakan keadaan menggunakan skor ujian atau pemarkahan untuk menentukan sama ada prestasi atau pencapaian calon adalah memenuhi atau melebihi sesuatu piawaian atau tahap prestasi minimum yang boleh diterima. Dalam keadaan ini, keputusan pencapaian setiap calon adalah tidak bergantung kepada prestasi pencapaian calon lain. Dalam konteks teori G, semua faset yang menyumbang kepada ralat dan juga interaksi semua faset dengan calon (objek pengukuran) dan

faset lain akan mempengaruhi magnitud skor calon dari segi keputusan mutlak (Swartz et al., 1999).

1.5.9 Keputusan relatif (*Relative decisions*)

Ia merujuk kepada keadaan menggunakan skor ujian atau pemarkahan untuk membuat keputusan berdasarkan kedudukan relatif atau susunan pangkatan calon dalam suatu kumpulan. Kedudukan relatif bagi setiap calon bukan sahaja ditentukan oleh prestasinya sendiri, tetapi juga oleh prestasi ahli-ahli lain dalam kumpulan perbandingan. Dalam konteks teori G, hanya interaksi faset ralat yang berkaitan dengan calon (objek pengukuran) akan menyebabkan perubahan kedudukan relatif calon (Swartz et al., 1999).

1.5.10 Model gabungan (*Mixed model*)

Semua model dalam teori G yang mempunyai satu atau lebih faset tetap di samping faset rawak (Shavelson & Webb, 1991, p. 66).

1.5.11 Model rawak (*Random model*)

Sesuatu model dianggap sebagai model rawak atau kesan rawak dalam teori G sekiranya semua faset pengukuran adalah faset rawak (Brennan, 2001, p. 96).

1.5.12 Objek pengukuran (*Object of measurement*)

Objek pengukuran atau populasi (calon dalam kes ini) biasanya dilambangkan dengan *p* (*persons*) adalah varian benar dalam sesuatu keadaan pengukuran (Suen, 1990). Berdasarkan tujuan kajian, *persons* atau calon, item, pemeriksa, keadaan, kelas dan sebagainya yang berkeadaan kumpulan boleh dianggap sebagai objek pengukuran.

1.5.13 Pekali *generalizability*, pekali G (*Generalizability coefficient, $E\rho^2$*)

Pekali G boleh ditakrifkan sebagai nisbah varian skor semesta kepada jangkaan varian skor cerapan (Brennan, 1992 & 2001; Cronbach et al., 1972; Shavelson & Webb, 1991). Secara mudah, ia boleh dianggar menerusi pembahagian varian sistematik dalam purata pemarkahan calon dengan hasil tambah varian sistematik dan varian ralat relatif. Ia menyerupai pekali kebolehpercayaan dalam teori ujian klasik (Brennan, 2001). Julat nilainya adalah di antara 0 dan 1. Ia boleh dilambangkan sebagai $E\rho^2$ dan rumusannya adalah seperti berikut:

$$E\rho^2 = \sigma^2(\tau) / [\sigma^2(\tau) + \sigma^2(\delta)],$$

di mana $\sigma^2(\tau)$ ialah varian skor semesta dan $\sigma^2(\delta)$ ialah varian ralat relatif.

1.5.14 Pekali phi atau Indeks kebergantungan (*Dependability Index, Φ*)

Indeks kebergantungan boleh ditakrifkan sebagai nisbah varian skor semesta bagi dirinya bercampur dengan varian ralat mutlak (Brennan, 2001, p.13). Secara mudah, ia boleh dianggar menerusi pembahagian varian sistematik dalam purata pemarkahan calon dengan hasil tambah varian sistematik dan varian ralat mutlak. Julat nilainya adalah di antara 0 dan 1. Ia boleh dilambangkan sebagai Φ dan rumusannya adalah seperti berikut:

$$\Phi = \sigma^2(\tau) / [\sigma^2(\tau) + \sigma^2(\Delta)],$$

di mana $\sigma^2(\tau)$ ialah varian skor semesta dan $\sigma^2(\Delta)$ ialah varian ralat mutlak.

1.5.15 Ralat mutlak (*Absolute Error, Δ*)

Ia ditakrifkan sebagai perbezaan antara skor cerapan dengan skor semesta bagi individu (Brennan, 2001, p.11).

1.5.16 Ralat relatif (*Relative Error, δ*)

Ia ditakrifkan sebagai perbezaan antara skor sisihan cerapan dengan skor sisihan semesta bagi individu (Brennan, 2001, p.12).

1.5.17 Reka bentuk tersilang (*Crossed design*)

Ia merupakan keadaan di mana objek pengukuran dapat diukur oleh semua paras yang dikenal pasti bagi semua faset pengukuran yang terlibat (Suen, 1990).

1.5.18 Reka bentuk tersarang (*Nested design*)

Ia merupakan keadaan di mana objek pengukuran hanya dapat diukur oleh sebahagian daripada semua paras yang dikenal pasti bagi satu atau lebih faset yang terlibat (Suen, 1990).

1.5.19 Reka bentuk separa tersarang (*Partially nested design*)

Reka bentuk ini adalah campuran daripada reka bentuk tersilang dan tersarang kerana ia mengandungi kedua-dua kesan tersilang dan tersarang (Shavelson & Webb, 1991, p. 52). Contohnya, reka bentuk dalam kajian ini iaitu $p \times (r{:}t)$ di mana $n_p = 120$ orang calon dan $n_t = 2$ tugasan dan setiap tugasan akan dinilai oleh kumpulan pemeriksa ($n_r = 3$) yang berbeza. Walaupun para pemeriksa tersarang dalam tugasan, namun kedua-dua faset pemeriksa dan tugasan adalah tersilang dengan calon iaitu setiap calon akan menjawab semua tugasan dan dinilai oleh kumpulan pemeriksa yang berlainan.

1.5.20 Semesta (*Universe*)

Konsep semesta dalam Teori G merujuk kepada keadaan di mana sesuatu populasi mungkin dicerap (Cronbach et. al., 1972, p. 9). Menurut Brennan (2001), istilah semesta sesuai untuk kondisi pengukuran seperti pemeriksa dan tugasan yang juga merupakan sumber variasi dalam pengukuran.

1.5.21 Semesta cerapan teraku (*Universe of admissible observations*)

Merujuk kepada cerapan-cerapan yang mana pembuat keputusan sudi menganggapnya sebagai sesuatu yang boleh saling bertukar ganti untuk tujuan membuat sesuatu keputusan dalam kajian G (Shavelson & Webb, 1991).

1.5.22 Semesta generalisasi (*Universe of generalization*)

Ia merupakan semesta di mana pembuat keputusan ingin membuat generalisasi dalam kajian D berdasarkan sesuatu prosedur pengukuran tertentu (Brennan, 2001). Suatu semesta generalisasi mungkin mengandungi semua kondisi dalam semesta cerapan teraku atau hanya mengandungi suatu subset mengenai kondisi dalam semesta cerapan teraku (Brennan, 1983, p. 3).

1.5.23 SJK(C)

Sekolah Jenis Kebangsaan Cina [SJK(C)] merupakan sekolah rendah di Malaysia yang menyediakan kanak-kanak dengan enam tahun pendidikan awal secara percuma. Guru menggunakan Bahasa Cina sebagai bahasa pengantar dalam

pengajaran kecuali mengajar mata pelajaran Agama Islam, Bahasa Inggeris dan Bahasa Malaysia. Namun, mulai tahun 2003, Kementerian Pendidikan Malaysia (KPM) telah menetapkan bahawa sebahagian daripada peruntukan masa pengajaran dan pembelajaran mata pelajaran Sains dan Matematik diajar dalam bahasa Inggeris (KPM, 2003). Sukatan dan kurikulum pelajaran SJK(C) adalah ditetapkan oleh KPM.

1.5.24 Skor semesta (*Universe score, μ*)

Skor semesta ditakrifkan sebagai min skor bagi calon berdasarkan semua cerapan yang boleh diterima dalam sesuatu semesta (Cronbach et. al., 1972, p. 15). Secara ringkas, skor semesta merupakan jangkaan skor (min skor) tercerap untuk suatu objek pengukuran dalam semesta generalisasi (Miller & Kane, 2001). Tujuan sesuatu pengukuran adalah untuk menganggar skor semesta dengan jitu berdasarkan sampel cerapan tertentu, yakni semesta cerapan teraku tertentu.

1.5.25 Teori *Generalizability* (Teori G)

Teori G merupakan perluasan metodologi dan teori kebolehpercayaan klasik yang menggunakan analisis varian untuk menganggar komponen varian. Teori G juga digunakan untuk menerangkan magnitud ralat daripada pelbagai sumber tertentu. Analisis tersebut digunakan untuk menilai pengitlakan skor yang melewati sampel item, calon dan kondisi cerapan yang ditentukan dalam kajian (AERA, APA & NCME, 1985).

1.5.26 Varian skor semesta

Ia merujuk kepada nilai jangkaan kovarian bagi min skor objek pengukuran ke atas pasangan kes-kes setara secara rawak dalam prosedur pengukuran (Brennan, 2001, p. 98) dan umumnya dilambangkan sebagai $\sigma^2(\tau)$.

1.5.27 Varian ralat mutlak (*Absolute error variance,* $\sigma^2\Delta$)

Ia merujuk kepada semua komponen varian kecuali varian yang disebabkan oleh faset yang diminati, dalam teori G dipanggil sebagai objek pengukuran dan dalam konteks kajian ini adalah calon, yang menyumbang kepada ralat pengukuran apabila pemarkahan digunakan sebagai asas untuk membuat keputusan mutlak. Oleh itu, varian ralat mutlak adalah hasil tambah semua komponen varian kecuali komponen varian bagi calon kerana calon tidak dianggap sebagai varian ralat (Shavelson & Webb, 1991, p. 84).

1.5.28 Varian ralat relatif (*Relative error variance,* $\sigma^2\delta$)

Hanya komponen-komponen varian yang mewakili interaksi dengan objek pengukuran (calon dalam kes ini) menyumbang kepada ralat pengukuran berkaitan dengan keputusan relatif (Shavelson & Webb, 1991: 84). Justeru itu, varian ralat relatif adalah hasil tambah semua komponen varian yang mewakili interaksi objek pengukuran dengan faset yang lain. Punca kuasadua untuk statistik ini adalah menyerupai ralat piawai pengukuran dalam teori ujian klasik.

1.6 Kepentingan Kajian

Kepentingan kajian ini dapat ditinjau dalam lima aspek. Pertama, dari sifat kajian ini sendiri. Walaupun terdapat kajian-kajian tertentu mengenai kemahiran

menulis dalam pentaksiran bahasa di negara ini (Abdul Aziz Abdul Talib, 1985; Awang Sariyan, 1987; Koh, 1985), namun kajian yang melibatkan penganalisisan dan penganggaran varian secara kuantitatif hasil interaksi faktor-faktor seperti karangan yang berbeza bentuk, pemeriksa dan prosedur pemarkahan secara serentak dengan menggunakan Teori *Generalizability* adalah buat pertama kali dijalankan.

Kedua, dari segi kepentingan pentakisran karangan. Dalam bidang pendidikan di mana-mana negara di dunia ini, kemahiran menulis adalah kemahiran yang paling penting dan asas (Kamarudin Hj. Husin, & Siti Hajar Hj. Abdul Aziz, 1997) yang perlu diajar kepada murid dari peringkat sekolah rendah lagi. Malah kemahiran menulis merupakan kayu ukur tanda celik huruf oleh masyarakat antarabangsa. Untuk mentaksir kemahiran menulis pelajar, penulisan karangan biasanya digunakan sama ada dalam pentaksiran berasaskan sekolah atau peperiksaan awam. Ini karana pakar pendidikan dan pentaksiran rata-rata menganggap kesahan ujian penulisan langsung adalah tinggi dan mempunyai impak positif dalam pengajaran daripada ujian objektif (Milanovic, Saville & Shen, 1993). Namun begitu, pentaksiran karangan sentiasa diancam oleh masalah keadilan dan ketekalan dalam pemberian skor menyebabkan skor yang diberi tidak dapat mencerminkan kemahiran menulis sebenar pelajar. Memandangkan hakikat ini, kajian ini bertujuan memberi gambaran yang lebih jelas dengan mengungkaikan faktor-faktor yang mempengaruhi kebolehpercayaan skor karangan menerusi analisis dan anggaran varian.

Sejajar dengan arus perkembangan pentaksiran portfolio (Yancey, 1999), pakar pentaksiran sedang mengkaji dan menerokai kebolehlaksanaan pengujian karangan ditaksir melalui portfolio. Masalah paling ketara yang dihadapi oleh

pentaksiran portfolio adalah pekali kebolehpercayaan yang rendah untuk tugasan yang ditaksir. Kajian ini meninjau tugasan karangan yang berbeza bentuk dinilai dengan aspek dan kaedah pemarkahan yang berlainan di samping memaparkan kombinasi tugasan dan pemeriksa dengan pekali G yang tinggi boleh memberi ilham dan asas rujukan kepada pihak sekolah atau badan pentaksiran untuk merangka strategi-strategi yang berkesan dalam mengawal ralat pemarkahan.

Keempat, dari segi maklumat dapatan kajian. Dapatan kajian ini menyampaikan maklumat kebolehpercayaan skor tentang aspek pemarkahan yang berlainan iaitu sama ada kandungan dan organisasi atau penggunaan bahasa dan mekanis dengan kaedah pemarkahan yang berlainan. Selain itu, maklumat mengenai kebolehan genaralisasi skor antara aspek dan kaedah pemarkahan yang berlainan juga dipaparkan. Semua maklumat ini boleh memanfaatkan guru bahasa di sekolah rendah terutamanya SJK(C) bukan sahaja dari segi pengajaran dan pembelajaran karangan tetapi juga dari segi pentaksiran karangan.

Selain itu, dapatan kajian ini memberi satu orientasi pemikiran baru dan meningkatkan kefahaman individu tentang kelebihan teori G menangani ralat pengukuran multi faset iaitu melibatkan karangan berbeza bentuk, pemeriksa, aspek pemarkahan dan kaedah pemarkahan yang berlainan secara terperinci terutamanya dalam pentaksiran karangan. Keadaan ini memang tidak dapat dijangkaui oleh teori ujian klasik. Dapatan kajian ini juga boleh dijadikan panduan dan rujukan kepada para penyelidik yang berminat untuk menerokai teori G berkaitan aspek-aspek pentaksiran bahasa yang lain khasnya dan bidang pentaksiran dan pengukuran amnya.

1.7 Limitasi Kajian

Memang tidak dapat dinafikan bahawa terdapat beberapa kekurangan dan kelemahan dalam kajian ini. Dapatan kajian ini terbatas kepada karangan bentuk naratif dan pendedahan yang dihasilkan oleh murid-murid Tahun Enam SJK(C) di sebuah daerah di Perak. Justeru itu, dapatan kajian ini tidak semestinya dapat digeneralisasikan kapada tugasan karangan yana lain dan semua murid Tahun Enam SJK(C) di Malaysia. Padahal, bentuk-bentuk karangan lain seperti pembahasan, deskriptif dan imaginatif mungkin memberi ralat skor yang berlainan dan mempengaruhi pencapaian mutu karangan murid tidak dapat ditunjukkan. Selain itu, penglibatan murid-murid daripada tahap-tahap lain dan kawasan lain mungkin membolehkan perbandingan yang lebih bermakna dilakukan dan dapat memberi gambaran yang lebih menyeluruh tentang maklumat kebergantungan skor dalam pentaksiran karangan juga tidak dapat dipaparkan.

Tumpuan kajian ini adalah untuk mengenal pasti sumber ralat pelbagai faset yang terdapat dalam pentaksiran karangan. Walaupun kajian ini telah menyenaraikan faset tugasan karangan yang berlainan, faset pemeriksa dan faset prosedur pemarkahan yang berlainan, namun masih terdapat juga faset-faset lain yang berpotensi bertindak sebagai sumber variasi seperti faset persekitaran, faset masa dan sebagainya yang boleh mempengaruhi skor karangan. Oleh kerana batasan dari segi tujuan kajian, faset-faset tersebut tidak diteliti.

Kepersisan dapatan kajian ini bukan sahaja bergantung pada tahap kefahaman murid terhadap tugasan karangan yang dikemukakan tetapi juga

kesungguhan mereka untuk memberi respons terhadap tugasan tersebut. Tugasan karangan dalam kajian ini digubal oleh Jawatankuasa Penggubalan Soalan Jabatan Pelajaran Perak berdasarkan spesifikasi tertentu dan melalui proses pemurnian berperingkat, namun minat, pengetahuan sedia ada dan latar belakang murid yang berbeza mungkin mempengaruhi mutu karangan yang disembahkan. Selain itu, ketekalan pemeriksa dalam memberi skor karangan mengikut rubrik dan kriteria pemarkahan serta tahap kefahaman mereka mengenainya adalah penting untuk mendapatkan data yang jitu. Walaupun pemeriksa telah menjalani latihan yang rapi, tetapi sifat subjektif manusia dan ciri individu adalah tidak mudah dikikis. Begitu juga kefahaman pemeriksa mengenai kriteria pemarkahan tertentu adalah berbeza sama ada berpunca daripada pengetahuan, komitmen atau kepakaran yang dimiliki.

Selain itu, peruntukkan masa iaitu 30 minit untuk menjawab setiap tugasan karangan mungkin menjejaskan dan tidak dapat menggambarkan kemahiran menulis sebenar calon. Karangan langsung berjangka (*timed impromptu essay*) sering dijadikan topik perbincangan (Alderson, 2002; Weigle, 2002) dari segi keaslian dan dikatakan memihak kepada penulis yang berupaya menulis dengan cepat di samping kekangan masa yang dihadapi oleh calon (Brown, 2004). Perkara ini akan dibincangkan selanjutnya dalam Bab 5.

Satu lagi kekurangan dalam kajian ini ialah pemarkahan karangan dibahagikan kepada dua aspek utama iaitu aspek kandungan dan organisasi serta penggunaan bahasa dan mekanis. Oleh itu, anggaran kebergantungan skor karangan hanya berasaskan dua aspek utama berkenaan. Walaupun pembahagian tersebut dapat memberi maklumat yang lebih menyeluruh dan selaras dengan keperluan

kajian, iaitu untuk meninjau perubahan kebergantungan skor karangan berlainan bentuk berdasarkan aspek dan kaedah pemarkahan yang berlainan, namun begitu, dapatan tentang perubahan kebergantungan skor untuk seluruh karangan tidak dapat diperolehi. Ini juga mengehadkan kebolehan generalisasi dapatan kajian bagi aspek tersebut.

1.8 Rumusan Bab

Di dalam bab ini telah dijelaskan beberapa perkara penting yang berkaitan dengan kajian. Perkara-perkara tersebut adalah masalah pengukuran mutu karangan yang melibatkan pelbagai faktor seperti pemeriksa, tugasan karangan, kaedah dan aspek pemarkahan yang berbeza. Faktor-faktor ini sama ada secara berasingan atau berinteraksi, menghasilkan sumber ralat pemarkahan dan ini akan menyukarkan interpretasi skor. Oleh itu, anggaran tentang kesan sumber ralat pemarkahan adalah penting untuk memberi pemahaman yang lebih mendalam tentang hubungan kompleks antara faktor-faktor tersebut dalam pengukuran mutu karangan agar mencapai kejituan pengukuran. Beberapa soalan kajian telah dikemukakan berdasarkan objektif kajian. Akhir sekali, perkara berkaitan dengan definisi istilah, kepentingan kajian dan limitasi kajian juga dijelaskan.

BAB II

KAJIAN-KAJIAN LITERATUR

2.0 Pengenalan

Perbincangan dalam bab ini lebih tertumpu kepada kajian-kajian lepas tentang faktor-faktor yang mempengaruhi kesan tugasan karangan, kesan pemeriksa, kesan prosedur pemarkahan yang melibatkan aspek pemarkahan dan kaedah pemarkahan, dan penggunaan teori G dalam pentaksiran karangan. Bab ini juga merangkumi kajian yang berkaitan dengan kerangka teori G dari segi teori dan penganggaran, perbandingan teori G dengan teori ujian klasik dari segi konsep asas dan konsep kebolehpercayaan.

2.1 Pendahuluan Faktor-Faktor Yang Mempengaruhi Kebolehpercayaan Atau Kebergantungan Skor Ujian Karangan

Teori G merupakan teori statistik mengenai kebergantungan pengukuran tingkah laku manusia (Shavelson & Webb, 1991). Dalam teori G, konsep kebergantungan boleh ditafsirkan sebagai sejauh mana skor ujian adalah tekal dan boleh dipercayai (*dependable*) untuk membuat sesuatu keputusan (Kunnan, 1992). Secara spesifik, kebergantungan suatu skor ujian boleh dilihat dari segi kepersisan generalisasi suatu ujian daripada skor cerapan kepada skor purata seseorang calon yang diperolehnya dalam semua kondisi pengukuran yang mungkin terdapat dan pembuat keputusan atau pengguna ujian sanggup menerimanya (shavelson & Webb, 1991, p.1). Ini bermakna persoalan kebolehpercayaan dalam teori G telah dilanjutkan

kepada kepersisan generalisasi atau kebolehan generalisasi skor (Cronbach, Gleser, Nanda, Rajaratnam, 1972, p. 15). Kebolehpercayaan skor dalam ujian adalah penting dan diperlukan kerana ciri tersebut merupakan prasyarat untuk menjamin kesahan prestasi calon (Thompson, 2003). Teras bagi kebolehpercayaan skor adalah untuk mengenal pasti varian ralat (Anastasi & Urbina, 1997).

Tujuan kajian ini adalah untuk mengkaji pengaruh dan hubungan faktor-faktor ralat pengukuran dalam pentaksiran karangan calon menerusi kemahiran menulis. Calon dikenal pasti sebagai varian benar dalam kajian ini sementara faktor tugasan karangan, faktor pemeriksa dan faktor prosedur pemarkahan merupakan varian ralat. Sebenarnya, dalam kajian kebolehpercayaan skor ujian karangan, dua faktor tersebut iaitu faktor pemarkahan (prosedur pemarkahan dan termasuk juga pemeriksa) dan faktor tugasan karangan merupakan penyumbang varian ralat utama yang sentiasa mendapat tumpuan dan perhatian khusus setiap masa (Bachman, Lynch, & Mason, 1995; Huot, 1990; Kroll, 1998; Schoonen, 2005).

2.1.1 Pengaruh Faktor Kesan Tugasan Karangan

Dalam pengertian yang luas, tugasan meliputi input calon termasuk juga tingkah laku calon dalam proses peperiksaan, tajuk karangan dan maklumat arahan. Namun dalam pengertian yang sempit, tugasan biasanya merujuk kepada jenis item ujian yang boleh mencungkil kebolehan kompleks dan kemahiran produktif secara langsung daripada calon (Davies, 2002).

Pada hakikatnya, pengertian tugasan dan item adalah berbeza. Walaupun fungsi mereka saling bertindih antara satu sama lain pada aspek-aspek tertentu

memandangkan kedua-duanya mencungkil maklumat tentang pencapaian kebolehan tertentu calon. Namun begitu, berbeza dengan item, tugasan biasanya berunsur prestasi. Selain itu, ruang lingkup tugasan adalah lebih luas dan begitu juga liputan dimensinya adalah lebih kompleks. Manakala istilah item selalunya digunakan dalam soalan objektif yang mempunyai lingkungan dan kandungan yang agak tertumpu. Justeru itu, karangan adalah sesuai digolongkan sebagai tugasan. Karangan sebagai pentaksiran prestasi (Madaus & O'Dwyer, 1999; Weigle, 2002) mengkehendaki calon membina atau membekalkan jawapan, membuat atau menghasilkan sesuatu untuk dinilai (Madaus & O'Dwyer, 1999). Calon menyampaikan idea, fakta, penjelasan, gambaran dan sebagainya yang berkaitan dengan pemikiran mereka ke dalam karangan berwadahkan bahasa. Oleh itu, usaha mengarang selalunya dikategorikan sebagai pemikiran tahap tinggi.

Faktor tugasan mungkin bertindak ke atas prestasi calon iaitu memberi kesan terhadap tugasan dan seterusnya mengakibatkan keberubahan skor. Bachman (1990) beranggapan bahawa ciri-ciri tugasan yang perlu disiapkan oleh calon merupakan salah satu faktor yang boleh mempengaruhi prestasi skor ujian calon. Banyak dapatan kajian empirikal telah melaporkan kebolehpercayaan skor yang rendah bagi tugasan karangan yang berlainan. Skor calon yang didapati daripada tugasan yang berlainan cenderung memperlihatkan korelasi yang rendah (Brown et al., 1991; Cantor & Hoover, 1986; Chen et al., 2007; Lehmann, 1990; Moss et al., 1982; Schoonen, 2005; Stevens & Clauser, 1996). Dalam perbincangan ini, perhatian akan diberikan kepada ciri-ciri dan sebab-sebab berlakunya kesan tugasan karangan kerana penekanan tersebut akan membantu pengkaji memahami dengan lebih mendalam lagi tentang persoalan ralat skor.

Kajian-kajian empirikal yang dijalankan oleh para penyelidik menunjukkan kesan tugasan mungkin dipengaruhi oleh faktor-faktor seperti bentuk karangan, tajuk dan rangsangan karangan. Para penyelidik berpendapat perbezaan yang dihasilkan oleh faktor-faktor tersebut boleh mempengaruhi tahap kesukaran tugasan karangan dalam takat tertentu dan seterusnya memberi kesan terhadap prestasi skor calon.

Banyak kajian telah dijalankan untuk meneliti kesan rangsangan tajuk terhadap prestasi penulisan pelajar. Secara umumnya, kajian-kajian tersebut melaporkan tidak terdapat perbezaan yang signifikan merentas pelbagai jenis tajuk karangan yang berlainan berhubung prestasi pelajar (Brossell,1983; Brossell & Ash, 1984; Greenberg, 1981; McAndrew, 1981). Dalam pada itu juga, kajian Hoetker dan Brossell (1989) dan Smith et al. (1985) pula menunjukkan bahawa terdapat kesan dari segi jenis dan rangsangan tajuk terhadap pemarkahan karangan tetapi kesannya tidak menyeluruh. Selain itu, kajian Schoonen (2005) ke atas tiga buah rangsangan karangan yang berlainan iaitu berunsur pemujukan, deskriptif dan arahan atau pendedahan yang ditulis oleh pelajar Gred 6 di Belanda tidak menunjukkan bahawa data bagi dua buah rangsangan karangan yang berunsur pendedahan adalah lebih menghampiri antara satu sama lain daripada rangsangan karangan berunsur pemujukan dan deskriptif. Walau bagaimanapun, Huot (1990, p. 245) berpendapat bahawa hasil kajian yang gagal menunjukkan terdapatnya hubungan yang jelas antara bentuk karangan dengan mutu penulisan (Greenberg, 1981; Quellmalz, Capell, & Chou, 1982; Reed, Burton, & Kelly, 1985) tidak boleh dijadikan alasan bahawa bentuk karangan tidak mempengaruhi penulisan karangan.

Terdapat kajian-kajian yang meneliti pengaruh bentuk karangan terhadap pentaksiran mutu karangan pelajar. Kegley (1986) menjalankan kajian ke atas empat bentuk karangan klasik iaitu naratif, deskriptif, pendedahan dan pemujukan. Peratus pelajar yang memperoleh keputusan penilaian yang memuaskan berdasarkan bentuk karangan adalah berbeza iaitu naratif (56%), deskriptif (43%), penerangan (41%), dan pujukan (31%). Beliau menyimpulkan bahawa bentuk karangan mempengaruhi pentaksiran kecekapan penulisan pelajar pada keseluruhannya. Sachse (1984) juga mendapat keputusan kajian yang sama tentang prestasi pelajar berdasarkan bentuk karangan yang berlainan. Dalam kajiannya, skor bagi pelajar Gred 5 dan 9 cenderung berubah-ubah mengikut bentuk karangan yang berbeza, dan karangan naratif berunsur ekspresif mendapat urutan kedudukan yang lebih tinggi berbanding dengan tugasan pemujukan.

Kajian Freedman and Pringle (1984) menunjukkan walaupun 98% daripada pelajar berumur 12 dan 13 boleh menghasilkan karangan naratif klasik dengan memuaskan, namun hanya 12.5% daripada mereka mahir dalam menulis karangan bentuk penghujahan. Manakala Prater and Padia (1983) mendapati pelajar Gred 4 dan 6 memperoleh pangkatan yang lebih tinggi dalam penulisan bentuk ekspresif daripada pemujukan dan pendedahan. Prater (1985) melaporkan keputusan yang sama untuk pelajar sekolah tinggi. Quellmalz et al. (1982) juga mendapati bahawa bentuk karangan mempengaruhi mutu penulisan pelajar sekolah tinggi yang berpencapaian sederhana dan tinggi. Berbeza dengan kajian-kajian lain, mereka berpendapat pelajar dalam kajian mereka mungkin kekurangan pengalaman untuk menangani tajuk naratif atau mungkin pemarkahan terhadap respons naratif adalah lebih tegas. Selain itu, Quellmalz et al. (1982) juga menyentuh tentang prestasi aspek

penulisan pelajar dan pengaruh bentuk karangan. Kajian mereka mendapati bentuk karangan yang berlainan boleh mempengaruhi aspek penilaian karangan iaitu aspek tanggapan keseluruhan, organisasi dan pemarkahan keseluruhan. Bagaimanapun, kajian mereka mendapati aspek mekanis, fokus dan sokongan tidak dipengaruhi oleh karangan yang berbeza bentuk.

Bentuk, tajuk dan rangsangan karangan yang berlainan berpotensi mempengaruhi tahap kesukaran tugasan karangan. Dalam mengulas pencapaian penulisan pelajar untuk tempoh tertentu iaitu dari tahun 1969 hingga 1979, laporan *National Assessment of Educational Progress* (NAEP, 1980) menyatakan bahawa pelajar berasa tahap kesukaran tugasan penulisan bentuk pemujukan yakni yang melibatkan penghujahan, adalah jauh lebih tinggi daripada tugasan naratif, deskriptif dan pendedahan. Dalam hubungan ini, McCann (1989) telah meminta seramai 95 orang pelajar Gred 6, 9 dan 12 menilai dan menulis karangan bentuk penghujahan. Hasil kajiannya menunjukkan pelajar menghadapi masalah untuk menghasilkan sesetengah ciri tertentu berkaitan dengan bentuk karangan tersebut. Dapatan kajian Liu dan Zhang (1998) tentang kebolehpercayaan skor pemarkahan karangan Bahasa Cina yang melibatkan pelajar sekolah menengah atas dan bahasa ibunda mereka ialah Bahasa Cina juga memperlihatkan bahawa ralat pemarkahan untuk karangan bentuk penghujahan adalah lebih besar daripada bentuk naratif serta gabungan bentuk penghujahan dan naratif.

Namun begitu, kajian Schoonen (2005) ke atas pelajar Gred 6 berumur 11 hingga 12 tahun yang menggunakan Bahasa Belanda sebagai bahasa ibunda pula memaparkan ralat pemarkahan bagi karangan berfungsi menerangkan dan memberi

arahan iaitu menyerupai karangan berunsur pendedahan adalah jauh lebih besar daripada karangan berunsur deskriptif dan pemujukan. Dalam kajian untuk melihat pengaruh tugasan penulisan yang berlainan bentuk terhadap kualiti dan kuantiti karangan bahasa asing bagi pelajar kolej di Amerika yang mengikuti program bahasa Jepun, dapatan kajian Koda (1993) menunjukkan bahawa tugasan berunsur naratif adalah lebih sukar daripada tugasan berunsur deskriptif kerana tugasan berunsur naratif mungkin melibatkan tuntutan pemprosesan linguistik yang lebih banyak pada tahap yang berbeza. Way, Joiner dan Seaman (2000) memperoleh keputusan kajian yang sama dalam satu kajian yang melibatkan 330 pelajar baru yang belajar Bahasa Perancis tentang kesan tugasan penulisan yang berlainan bentuk. Dapatan kajiannya menunjukkan bahawa tugasan bentuk pendedahan adalah paling sukar, diikuti oleh tugasan bentuk naratif dan yang paling mudah ialah bentuk deskriptif. Walau bagaimanapun, hasil kajian Scott (1996) tentang penulisan bentuk naratif dan deskriptif dalam bahasa pertama dan kedua adalah bertentangan. Dalam kajiannya, Scott mendapati karangan bentuk naratif adalah lebih mudah daripada bentuk deskriptif manakala tahap kesukaran karangan bentuk pendedahan dalam kajian bahasa pertama dan kedua adalah selari, yakni bentuk pendedahan adalah lebih sukar daripada bentuk naratif dan deskriptif.

Kebanyakan dapatan kajian yang telah dibincangkan memaparkan hubungan statistik antara bentuk, tajuk dan rangsangan karangan dengan prestasi penulisan calon hasil daripada kajian empirikal, namun kajian-kajian tersebut masih kurang menyeluruh untuk menjelaskan pengaruh bentuk, tajuk dan rangsangan karangan terhadap mutu penulisan pelajar.

Sejak sepuluh tahun kebelakangan ini, perkembangan psikologi kognitif dalam bidang linguistik gunaan telah mendorong para penyelidik mencari penjelasan untuk persoalan tersebut dari perspektif sains kognitif. Misalnya, ada penyelidik berpendapat bahawa kebiasaan (*familiarity*) calon tentang sesuatu bentuk karangan akan mempengaruhi jangkaan tahap kesukaran keseluruhan tugasan tersebut (Franken & Haslett, 2002). Kajian lepas tentang bentuk karangan memperlihatkan bahawa pelajar menunjukkan prestasi yang lebih baik terhadap tugasan penulisan yang biasa kepada mereka. Justeru itu, kebiasaan atau pengetahuan sedia ada pelajar mungkin merupakan satu jangkaan yang ketara bagi mutu penulisan (Britton et al., 1975; Kinneavy, 1971; Quellmalz, 1984).

Kebiasaan calon terhadap bentuk karangan tertentu mungkin menggerakkan struktur kebolehan dalaman calon untuk menghadapi tugasan yang diberikan. Menurut Kellogg (2007), kebiasaan merupakan proses automatik yang boleh mempengaruhi ingatan seseorang tanpa disedari. Sebaliknya, calon yang tidak biasa dengan bentuk karangan tertentu berasa tugasnya lebih sukar dan ini akan menambahkan lagi bebanan kognitif (Foster & Skehan, 1996; Sweller, 1994). Bebanan tersebut akan menyebabkan calon menumpukan perhatian dalam proses penelitian (*reviewing*), penjanaan (*generating*) dan perancangan (*planning*) serta membuat penyesuaian secara berterusan antara ketiga-tiga proses tersebut semasa mengarang (Sweigart, 1991).

Walaupun skema kognitif mencerminkan bahawa bentuk karangan yang berlainan memerlukan jenis pengetahuan yang tidak sama, namun dalam usaha mengarang, penulis juga perlu mengeksploitasi pengetahuan yang berkaitan dengan

kandungan atau maklumat tambahan. Berdasarkan kajian rintis Greenberg (1981), tuntutan pengalaman (*experiential demand*) yang diperlukan dalam tugasan penulisan adalah berbeza iaitu merangkumi respons yang menuntut pengalaman personal penulis hinggalah pengetahuan tentang fakta dan generalisasi. Namun begitu, Brossell dan Ash (1984), Greenberg (1981), serta Hoetker dan Brossell (1989) mendapati pengaruh tuntutan pengalaman dalam pentaksiran mutu penulisan pelajar adalah tidak ketara.

Terdapat penyelidik yang cuba menghubungkaitkan minat individu dengan bentuk karangan bagi menjelaskan pengaruh bentuk karangan terhadap prestasi calon yakni kecenderungan calon terhadap bentuk karangan tertentu akan mempengaruhi prestasi mutu penulisannya. Namun, hipotesis ini masih tidak dapat dibuktikan. Kajian secara sistematik oleh Hidi dan McLaren (1990) terhadap karangan bentuk pendedahan didapati tidak menemui sebarang pertalian antara bentuk karangan dengan minat individu. Dalam hal yang sama, kajian Hidi dan Anderson (1992) juga menyatakan bahawa para penyelidik tidak menemui peranan penting yang dimainkan oleh pengaruh faktor minat calon terhadap bentuk karangan.

Lazimnya, semasa menduduki ujian penulisan terutamanya ujian yang dianggap oleh calon sebagai ujian berkepentingan tinggi (*high-stake test*), calon akan berusaha sedaya upaya untuk memberi respons tanpa mengira sama ada mereka berminat ataupun mempunyai pengalaman tentang tajuk karangan tersebut (Powers & Fowles, 1999). Walaupun fenomena tersebut wujud secara meluas, namun para penyelidik masih meneruskan usaha untuk mengkaji sama ada terdapat pertalian antara prestasi penulisan dengan minat atau pengetahuan individu terhadap tajuk

karangan tertentu (Benton et al., 1995). Mereka mendapati minat atau pengetahuan yang dimiliki calon terhadap sesuatu tajuk karangan boleh mempengaruhi prestasi penulisan calon terutamanya aspek kandungan dan organisasi mengalami pengaruh yang paling besar (Schoonen, 2005). Hidi dan Anderson (1992) pula menemui satu perkaitan yang unik dan kompleks antara minat dan pengetahuan calon terhadap sesuatu tajuk karangan dengan prestasi penulisan mereka. Minat calon terhadap sesuatu tajuk karangan tidak dapat memampasi kekurangan pengetahuan tentang tajuk karangan tertentu, sebaliknya memiliki pengetahuan yang cukup mantap tentang sesuatu tajuk boleh memampasi kekurangan minat terhadap sesuatu tajuk karangan. Manakala Tobias (1994) membuat kesimpulan bahawa pengetahuan yang dimiliki oleh calon dan minat mereka terhadap sesuatu tajuk karangan mempunyai hubungan linear yang sangat kuat.

Kellogg (1987) beranggapan bahawa sekiranya calon memiliki semakin banyak pengetahuan tentang sesuatu tajuk, semakin banyak usaha akan dicurahkan untuk merancang dan membina idea. Walau bagaimanapun, sekiranya keadaan sebalik yang berlaku, usaha akan dialihkan kepada proses penelitian dan pentafsiran (*translating*). Kajian beliau juga menunjukkan penulis yang berpengetahuan memperuntukkan usaha kognitif yang lebih banyak berbanding dengan penulis yang kurang berpengetahuan dalam proses penulisan. Ini mungkin kerana minat calon terhadap sesuatu tajuk karangan akan mendorong mereka mencurahkan lebih banyak sumber dan usaha daripada tajuk yang tidak diminati oleh mereka (Hidi, 1990) agar dapat mencapai prestasi penulisan yang unggul (Kellogg, 1987). Selain itu, pengalaman lalu individu juga berpotensi mempengaruhi prestasinya terhadap sesuatu tugasan. Flower (1994) beranggapan bahawa pengalaman lalu yang dimiliki

oleh seseorang boleh mempengaruhi pemahamannya tentang sesuatu tugasan dan penyusunan strategi dalam penulisannya. Menurut Langer (1984), latar belakang pengetahuan yang tersusun rapi mengenai sesuatu tajuk akan memanfaatkan penulisan seseorang.

Pada hakikatnya, adalah tidak mungkin untuk menghapuskan kesan tugasan secara menyeluruh sekiranya kita benar-benar memahami pengertian tentang reka bentuk, pembangunan (*development*) dan penggunaan ujian. Oleh itu, inisiatif perlu diambil untuk memahami dan mengawal kesan tugasan karangan supaya kesan tersebut dapat diminimumkan melalui penyediaan dan penggunaan soalan ujian yang sesuai dan bermutu tinggi (Bachman & Palmer, 1996).

2.1.2 Pengaruh Faktor Kesan Pemeriksa

Pemeriksa adalah salah satu komponen penting dalam proses pemarkahan. Biasanya skrip jawapan karangan akan dinilai oleh sekumpulan pemeriksa berpandukan prosedur pemarkahan yang ditetapkan. Prosedur pemarkahan bertujuan memandu pemeriksa menilai skrip jawapan dengan memberi markah secara adil dan terselaras. Secara tabii, pembinaan item ujian subjektif adalah lebih mudah berbanding dengan item ujian objektif. Namun begitu, kerja pemarkahan bagi ujian subjektif adalah lebih rumit. Kerumitan ini timbul kerana pemarkahan melibatkan aspek-aspek yang berkaitan dengan faktor manusia dan faktor prosedur.

Sejak akhir abad ke-19 lagi, kajian tentang keberubahan skor ujian soalan subjektif yang berkaitan dengan faktor pemeriksa merupakan kajian yang penting dalam bidang pengujian (Lumley & McNamara, 1995). Lazimnya, usaha dan ikhtiar

diambil untuk mengurangkan keberubahan skor yang disebabkan oleh tingkah laku pemeriksa dalam pentaksiran prestasi. Ruch seawal tahun 1929 lagi sudah menyuarakan bahawa kesan pemarkahan yang subjektif mungkin boleh dikurangkan separuh seandainya pemeriksa menerima dan mematuhi peraturan pemarkahan yang disediakan dalam peperiksaan esei (Linacre, 1991, dalam Ruch, 1929).

Ketekalan pemberian markah oleh pemeriksa biasanya dibahagikan kepada dua jenis iaitu kebolehpercayaan intra pemeriksa dan kebolehpercayaan antara pemeriksa. Jenis kedua merujuk kepada darjah ketekalan bagi dua atau lebih pemeriksa yang menilai prestasi calon (Bachman, 1990). Jacobs (1981) dan Huot (1990) pernah menyentuh tentang eksperimen yang dijalankan oleh Diederich, French dan Carlton pada tahun 1961 iaitu seramai 53 orang pemeriksa telah menilai 300 buah karangan secara berasingan dengan skala 9 mata berdasarkan piawaian masing-masing dan hasil dapatannya menunjukkan setiap karangan menerima lebih daripada 5 jenis pemarkahan yang berbeza. Cason dan Cason (1984) juga mendapati bahawa peratus varian yang berpunca daripada perbezaan ketegasan antara pemeriksa dan perbezaan antara kebolehan calon adalah lebih kurang sama (35% dan 40%). Namun begitu, berikutan dengan kajian yang lebih lanjut, penyelidik bukan sahaja cuba mencari cara yang sesuai untuk mengurangkan perbezaan antara pemeriksa tetapi juga menerokai punca yang mempengaruhi perbezaan antara pemeriksa. Kajian-kajian tersebut pada asasnya melibatkan aspek-aspek berikut: kajian tentang latihan pemeriksa, kajian tentang ciri-ciri perbezaan latar belakang pemeriksa dan ketekalan pemarkahan, kajian tentang aspek psikologi pemeriksa dalam proses pemeriksaan, penggolongan pemeriksa berdasarkan kriteria pemarkahan dan sebagainya.

Rata-rata dapatan kajian menunjukkan bahawa latihan pemeriksa boleh mendatangkan kesan positif terhadap pemeriksa. Kesimpulan ini mempunyai asas bukti empirikal tertentu (Moon & Hughes, 2002; Weigle, 1994; Weigle, 1998). Kajian Lumley dan McNamara (1995) menunjukkan bahawa latihan pemeriksa adalah satu proses berulangan dan berterusan yang melibatkan sesi moderasi kerana kesan latihan mungkin tidak akan berkekalan lama. Walaupun terdapat kajian yang menunjukkan bahawa latihan pemeriksa barangkali boleh mengurangkan perbezaan markah yang ekstrim, namun persoalan perbezaan dan ketekalan pemeriksa masih tidak boleh dipandang sepi (Lunz, Wright, & Linacre, 1990). Menurut Lumley dan McNamara (1995) sumbangan utama pengendalian latihan pemeriksa adalah dari segi mengurangkan ralat rawak dalam proses penilaian.

Dapatan kajian Weigle (1998) tentang perbezaan ketegasan dan ketekalan pemeriksa, antara pemeriksa berpengalaman dengan yang tidak berpengalaman sebelum dan selepas menjalani latihan juga menunjukkan bahawa latihan pemeriksa adalah lebih berjaya membantu pemeriksa dari segi pemberian skor ramalan (*predictable scores*) iaitu kebolehpercayaan intra pemeriksa daripada pemberian skor persamaan (*identical scores*) yakni kebolehpercayaan antara pemeriksa. Ini adalah sejajar dengan pendapat Wiseman (1949) dan Wigglesworth (1994) iaitu latihan pemeriksa menjadikan pemeriksa bersifat lebih tekal kendiri (*self-consistent*). Kajian Miller dan Legg (1993) juga mendapati bahawa perbezaan aras kesukaran tugasan penulisan peperiksaan boleh menimbulkan perbezaan pemberian skor dalam kumpulan pemeriksa tertentu. Selain itu, penyelidik juga berusaha melihat pengaruh perbezaan latar belakang pemeriksa terhadap pemarkahan karangan. Dalam meninjau faktor latar belakang pemeriksa, terdapat penyelidik yang memilih perspektif

berlainan, ada kajian yang berdasarkan latar belakang budaya pemeriksa yang berbeza (Kobayashi & Rinnert, 1996; Shi, 2001 & 2003; Song & Caruso, 1996), ada kajian pula berdasarkan ciri perbezaan skor dari segi penggolongan pemeriksa kepada pemeriksa pakar dan pemeriksa biasa (Barnwell, 1989; Cumming, 1990; Shohamy, Gordon, & Kraemer, 1992; Schoonen, Vergeer & Eiting, 1997). Aspek psikologi pemeriksa juga boleh menghasilkan perbezaan pemberian skor terhadap calon dan kriteria pemarkahan. Dalam kajian Kondo-Brown (2002) tentang cara penilaian pemeriksa terlatih yang bersikap prasangka terhadap jenis calon dan kriteria tertentu dalam mentaksir penulisan bahasa Jepun sebagai bahasa kedua. Dapatan kajiannya menunjukkan bahawa pemeriksa mengamalkan ketegasan yang berbeza dalam menilai calon dan kriteria tertentu, dan pola prasangka bagi setiap pemeriksa pula adalah berlainan.

Kajian mutakhir yang dijalankan oleh Eckes (2008) ke atas 64 pemeriksa berpengalaman yang pernah terlibat dalam pemarkahan penulisan skala besar untuk menggolongkan pemeriksa mengikut kriteria pemarkahan tertentu. Pemeriksa diminta menunjukkan tahap kepentingan terhadap sembilan kriteria pemarkahan yang merangkumi kelancaran, kesempurnaan dan ketepatan tatabahasa yang biasa digunakan dalam pemarkahan berdasarkan skala empat mata. Kajian awalnya menunjukkan pandangan pemeriksa berbeza secara signifikan terhadap kriteria pemarkahan yang berlainan. Dapatan kajiannya menunjukkan pemeriksa boleh digolongkan kepada kategori-kategori tertentu berdasarkan profil kriteria pemarkahan yang berbeza. Kajian juga mendapati penggolongan tersebut sebahagiannya boleh dikaitkan dengan latar belakang pemeriksa. Kesimpulannya

adalah pemeriksa tidak dapat bertindak adil atau mengagihkan perhatian yang sama rata ke atas semua kriteria pemarkahan yang diberikan.

2.1.3 Pengaruh Faktor Kesan Prosedur Pemarkahan

Dalam kajian ini, prosedur pemarkahan meliputi aspek dan kaedah pemarkahan. Aspek pemarkahan merangkumi aspek kandungan dan organisasi serta aspek penggunaan bahasa dan mekanis. Manakala kaedah pemarkahan yang dikaji meliputi kaedah holistik dan analitik. Dalam pentaksiran formal yang bercirikan pemiawaian, setiap pemeriksa perlu mengikuti prosedur pemarkahan secara ketat dan terperinci semasa memberi skor atau gred. Kaedah pemarkahan yang biasa digunakan ialah analitik dan holistik (Linn & Ground, 2000). Kedua-dua kaedah ini mendapat sambutan yang meluas dalam bidang pengajaran dan pentaksiran bahasa.

2.1.3.1 Kaedah Pemarkahan Holistik

Kaedah pemarkahan holistik digunakan untuk menilai pencapaian keseluruhan penulisan calon dengan memberi satu skor tunggal sahaja berpandukan tanggapan pemeriksa ke atas kualiti tugasan calon mengikut satu set kriteria (Arshad Abd. Samad, 2004). Kriteria-kriteria itu mencakupi semua jangkaan tentang produk tersebut atau kandungan piawai yang disasarkan. Elliot, Plata dan Zelhart (1990) menggambarkan pemarkahan holistik seperti berikut:

> *Melihat sampel penulisan secara holistik adalah cuba melihat penulisan tersebut lebih daripada hanya mencampurkan bahagian-bahagian asasnya. Dalam mempertimbangkan sampel penulisan dari perspektif holistik, pembaca bukan menilai faktor-faktor khusus secara berasingan seperti pengendalian topik, cara pemilihan retorik dan kata, dan selok-belok sampingan yang membentuk hasil*

> *penulisan tersebut. Walaupun begitu, pemeriksa diminta menimbangkan faktor-faktor tersebut sebagai satu elemen yang boleh menyumbang ke arah penghasilan satu tanggapan keseluruhan kepada pembaca. Tanggapan keseluruhan inilah yang dicari dalam pemarkahan holistik.* (p. 17)

Memandangkan prestasi calon dinilai secara keseluruhan, maka pemeriksa tidak mungkin menghukum calon tentang aspek penulisan yang lemah (Cohen, 1994), sebaliknya menekankan kelebihan calon daripada kelemahannya (White, 1985). Ini juga menyebabkan kaedah pemarkahan ini didakwa mengabaikan kelemahan aspek penulisan pelajar (Cumming, 1990; Hamp-Lyons, 1990; Reid, 1993; Elbow, 1999). Sebenarnya kaedah pemarkahan seakan-akan kaedah ini telah tersebar secara meluas pada tahun 1960an (Godshalk, Swineford, & Coffman, 1966). Kelebihan pemarkahan holistik membolehkan pemeriksa melakukan penilaian berdasarkan tanggapan keseluruhan terhadap mutu karangan calon yang menepati tahap kebolehan yang digambarkan dalam setiap aspek pemarkahan.

Cara seperti ini boleh meningkatkan kecekapan pemeriksa dalam pemberian skor. Namun, ramai juga menyuarakan pendapat yang bertentangan kerana kaedah ini meletakkan sifat kebolehan bahasa yang sememangnya kompleks lagi berbeza dalam satu skala yang sama untuk dinilai secara keseluruhan. Keadaan ini boleh menyukarkan pemeriksa menerangkan dan menjustifikasikan keputusan yang dibuat. Biarpun mereka cuba membuat penjelasan tetapi adalah kabur dan kurang meyakinkan (Purves, 1992). Cumming, Kantor dan Power (2002) menyatakan bahawa cara pemarkahan yang mudah dan kriteria pemarkahan yang kurang jelas telah menjejaskan kelebihan yang dimiliki oleh kaedah ini. Cohen (1994, p. 315)

telah menyenaraikan sebanyak sepuluh kelemahan tentang pemarkahan holistik yang wujud dalam pentaksiran penulisan bahasa kedua.

Pemarkahan holistik dikatakan berkesan apabila kandungan piawaian yang ditaksir adalah saling berkaitan antara satu sama lain yang memudahkannya diskor sebagai satu entiti. Menurut Hughes (1989), ciri-ciri tersendiri sesuatu teks seperti tatabahasa, ejaan dan organisasi tidak sepatutnya dianggap sebagai satu entiti yang berasingan. Cara pemarkahan ini juga didapati lebih cepat, menjimatkan masa dan kos efisien. Oleh itu, kaedah ini dianggap sesuai diamalkan dalam peperiksaan skala besar yang melibatkan jumlah calon yang ramai (Bauer, 1981) dan juga tempoh masa pemberitahuan keputusan peperiksaan yang singkat atau pada masa yang ditetapkan. Kebolehpercayaan kaedah ini lebih terjamin sekiranya pemeriksa menjalani sesi latihan pemarkahan sampel skrip karangan agar satu piawaian dipegang bersama. Namun , kaedah ini selalu disalahgunakan kerana pengguna terlalu menekankan pencapaian produk akhir sehingga mengabaikan proses pembelajaran pelajar.

2.1.3.2 Kaedah Pemarkahan Analitik

Pemarkahan analitik adalah jenis pemarkahan yang bertentangan dengan pemarkahan holistik. Ia merujuk kepada satu set kriteria yang digunakan untuk menilai faktor-faktor istimewa berkaitan dengan keupayaan calon dalam memaparkan kecekapan tentang sesuatu jangkaan yang spesifik atau kandungan piawai hasil daripada pengajaran dan pembelajaran. Setiap jangkaan tersebut ditaksir dengan satu panduan atau komponen panduan pemarkahan yang berasingan. Biasanya panduan pemarkahan analitik ini dibina dengan bentuk matriks. Dalam matriks tersebut, setiap jangkaan mempunyai satu set kriteria atau bahagian-bahagian

asasnya yang tersendiri dan boleh diskor secara berasingan. Setelah tugasan dinilai, skor-skor berasingan ini boleh dicampurkan dan memberikan satu skor muktamad.

Kaedah pemarkahan ini memberi kesedaran kepada pemeriksa tentang pemarkahan holistik yang mungkin mengabaikan aspek-aspek tertentu. Pemarkahan ini adalah lebih tertumpu dan faktor subjektif mungkin berkurangan. Namun begitu, Hughes (1989) berpendapat penumpuan ke atas aspek-aspek yang berlainan akan mengalihkan perhatian pemeriksa terhadap tanggapan penilaian keseluruhan penulisan esei tersebut. Tambahan pula, kebanyakan skema pemarkahan analitik memperuntukkan wajaran yang berlainan kepada aspek-aspek tertentu berdasarkan kepentingan sesuatu aspek. Ini akan merumitkan lagi pemarkahan.

Cara pemarkahan analitik dapat mendedahkan kemahiran yang belum dikuasai oleh pelajar. Oleh itu, cara pemarkahan ini dapat memberi maklum balas yang bermanfaat tentang prestasi pelajar untuk setiap elemen atau kriteria yang ditaksir (Roid, 1994). Kebanyakan guru mendapati kaedah ini adalah sangat sesuai untuk situasi pembelajaran dan pengajaran dalam kelas kerana mereka boleh mengesan aspek-aspek kekuatan dan kelemahan penulisan pelajar agar memudahkan penyusunan strategi pembelajaran dan pengajaran. Semasa mempraktikkan pemarkahan analitik, pemeriksa perlu memberi pertimbangan kepada aspek-aspek pencapaian pelajar agar tidak mengabaikan mana-mana satu aspek yang ditaksir. Keadaan ini dikatakan menjadikan pemarkahan tersebut mempunyai kebolehpercayaan yang tinggi. Dapatan kajian Huot (1990b) menunjukkan bahawa kebolehpercayaan pemarkahan analitik adalah lebih tinggi daripada pemarkahan holistik. Bagaimanapun, penggunaan kaedah ini memakan masa dan kos

pentadbirannya tinggi dan cenderung diganggu oleh kesan *Halo* semasa pemarkahan. Justeru itu, ia harus digunakan secara berhati-hati. Secara ringkas, ciri-ciri pemarkahan analitik dan holistik adalah seperti dalam Jadual 2.1.

Jadual 2.1

Perbandingan Ciri-Ciri Pemarkahan Kaedah Analitik Dan Kaedah Holistik

Bil	Pemarkahan Analitik	Pemarkahan Hoslitik
1.	Skor yang berasingan untuk setiap satu kriteria	Satu skor tunggal berdasarkan pelbagai kriteria
2.	Mengambil masa yang lebih panjang	Mudah dan menjimatkan masa
3.	Berkesan dan sesuai untuk mendiagnosis kemajuan dan keperluan individu	Berkesan dan sesuai untuk pentaksiran peringkat akhir atau sumatif
4.	Dapat menilai kandungan standard yang sangat spesifik	Dapat menilai hasil penulisan sebagai satu entiti
5.	Pemarkahan berdasarkan bahagian yang berasingan yang membentuk penulisan tersebut	Pemarkahan berdasarkan tanggapan keseluruhan (*total impression*)

2.1.3.3 Aspek-Aspek Pemarkahan

Aspek pemarkahan dalam pengujian dan pengukuran karangan merupakan pengenalpastian dan penggolongan ciri-ciri tahap kebolehan menulis pelajar. Apa sekalipun aspek pemarkahan dan kaedah pemarkahan yang digunakan, pada akhirnya kerja pemeriksaan terpaksa juga dilakukan oleh pemeriksa. Coffman (1971) membincangkan tiga jenis variasi pemarkahan antara pemeriksa, iaitu (a) perbezaan ketegasan pemberian markah antara pemeriksa ekoran daripada perbezaan piawaian yang digunakan masing-masing; (b) pemeriksa mempunyai kecenderungan berbeza dalam pengagihan markah sepanjang skala pemarkahan; dan (c) perbezaan pemahaman pemeriksa tentang kriteria pemarkahan menyebabkan mereka memberi markah yang berbeza terhadap karangan yang sama. Ini jelas menunjukkan bahawa

punca perbezaan pemarkahan pemeriksa adalah berkait rapat dengan tahap pemahaman pemeriksa terhadap kandungan yang diterangkan dalam aspek pemarkahan. Oleh itu, pemantapan piawaian pemarkahan adalah penting dalam mengurangkan perbezaan skor antara pemeriksa (Nitko, 2004).

Schoonen et al. (1997) telah menjalankan kajian tentang pemarkahan karangan yang dilakukan oleh pemeriksa pakar dan pemeriksa biasa ke atas sampel penulisan pelajar Gred 6 (min umur 12 tahun). Dapatan kajian mereka tentang tugasan spesifik yang digubal dalam tiga topik tertentu memaparkan nilai kebolehpercayaan antara pemeriksa bagi pemeriksa biasa yang kurang berpengalaman adalah lebih rendah dalam pemarkahan aspek penggunaan bahasa iaitu termasuklah kosa kata, tatabahasa dan gaya (.30, .70 dan .61). Manakala dalam aspek kandungan bahasa pula, nilai kebolehpercayaan antara pemeriksa adalah lebih memuaskan (.90, .70 dan .91). Selain itu, perbezaan skor antara pemeriksa pakar dan pemeriksa biasa adalah lebih kecil dalam aspek kandungan daripada aspek penggunaan bahasa. Ini mencerminkan bahawa dalam takat tertentu, aspek pemarkahan mempunyai fungsi kawalan ke atas kesan pemeriksa iaitu indeks kebolehpercayaan pemeriksa akan berubah dengan berubahnya pemberian aspek pemarkahan yang berlainan (sama ada kandungan atau penggunaan bahasa).

Faktor prosedur pemarkahan selain mempengaruhi pemeriksa dalam takat tertentu, mungkin juga mengenakan fungsi kawalan tertentu terhadap kesan faktor tugasan penulisan. Dalam mengkaji hubungan berkaitan dengan skor merentas tiga jenis tugasan pentaksiran (pentaksiran langsung, sampel naratif bilik darjah dan koleksi naratif) dengan dua jenis rubrik pemarkahan penulisan naratif kaedah holistik

iaitu satu rekaan baru dan satu lagi sedia ada sebagai perbandingan, Novak, Herman dan Gearhart (1996) cuba meneliti dua keadaan iaitu tugasan pentaksiran yang sama dinilai oleh rubrik pemarkahan yang berbeza untuk mendapat bukti kesahan *divergent*, dan tugasan pentaksiran yang berlainan dinilai oleh rubrik pemarkahan yang sama untuk mendapat bukti kesahan *convergent* bagi mengesahkan hipotesis yang sebelumnya. Namun, hubungan tersebut adalah sukar diwujudkan dalam dapatan kajian mereka (Banding dengan Quellmaz et al., 1982). Ini mungkin disebabkan kesan tugasan penulisan telah mengkaburkan rubrik pentaksiran yang ditetapkan dalam aspek pemarkahan dan menimbulkan perbezaan antara rubrik pentaksiran dengan aspek pemarkahan. Oleh itu, ciri-ciri tugasan penulisan seperti tajuk dan bentuk karangan mungkin mengenakan kesan yang lebih besar daripada skor sebenar kebolehan menulis calon. Sebaliknya, sama ada besar atau kecilnya kesan sesuatu tugasan penulisan adalah bersandar kepada rubrik yang berlainan yang terdapat dalam aspek pemarkahan (Novak, Herman & Gearhart, 1996, p. 232).

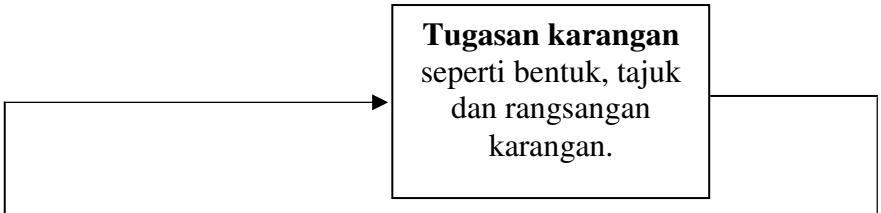
Dalam kajian Schoonen (2005), aspek pemarkahan yang digunakan hanya dua iaitu aspek kandungan dan organisasi serta aspek penggunaan bahasa. Hipotesisnya menyatakan aspek pertama lebih bersandar kepada ciri tugasan penulisan berbanding dengan aspek kedua. Menurut beliau, ini mungkin kerana pengaruh pengetahuan tentang tugasan penulisan yang dimiliki pelajar mempunyai kesan yang lebih kuat terhadap prestasi pelajar dalam aspek kandungan dan organisasi. Manakala kemahiran penggunaan bahasa pelajar kurang bergantung pada ciri tugasan. Kajiannya membuktikan aspek pemarkahan berfungsi sebagai pengantara terhadap kesan tugasan penulisan. Kajian Quellmaz et al. (1982) juga melaporkan bahawa pengaruh aspek organisasi bagi karangan bentuk naratif dan

pendedahan pelajar Gred 11 dan 12 adalah lebih besar daripada pengaruh aspek penggunaan bahasa.

Berdasarkan kajian empirikal yang dibincangkan, adalah wajar untuk membuat jangkaan bahawa dalam proses pemarkahan ujian karangan, prosedur pemarkahan mungkin bertindak dengan faktor pemeriksa dan tugasan penulisan pada masa yang serentak, seterusnya menghasilkan kesan tertentu antara mereka dan menyebabkan berlakunya keberubahan skor. Hipotesis kajian Schoonen (2005) menyatakan bahawa kesan pemeriksa dan tugasan penulisan mungkin banyak dipengaruhi oleh kaedah dan aspek pemarkahan. Dengan kata lain, kaedah dan aspek pemarkahan yang berlainan mungkin berupaya bertindak sebagai pengantara terhadap kesan pemeriksa dan tugasan penulisan. Interaksi keempat-empat faktor iaitu kesan pemeriksa, tugasan penulisan, kaedah pemarkahan dan aspek pemarkahan mempunyai perkaitan yang amat kompleks. Hasil eksperimen tentang kerumitan hubungan antara tugasan penulisan, pemeriksa dengan prosedur pemarkahan membuktikan jalan pemikirannya iaitu jumlah varian bagi kesan tugasan penulisan dan pemeriksa berdasarkan kaedah pemarkahan yang berlainan (holistik dan analitik) dan aspek pemarkahan yang berlainan (kandungan dan penggunaan bahasa) mempunyai makna perbezaan yang tertentu. Menurut Schoonen (2005) lagi, dengan menggunakan pemarkahan holistik, pertambahan varian kesan tugasan penulisan berbanding dengan varian keseluruhan adalah tidak begitu nyata tetapi dengan penggunaan pemarkahan analitik pertambahannya adalah jelas. Selain itu, perbezaan pekali generalizabiliti bagi kesan pemeriksa dan tugasan penulisan berdasarkan kaedah dan aspek pemarkahan yang berlainan adalah agak besar. Misalnya, pekali G bagi kandungan dan organisasi adalah paling rendah (.21) apabila kaedah analitik digunakan untuk menilai sampel empat jenis karangan yakni tajuk dan bentuk

karangan adalah berlainan. Manakala pekali G bagi penggunaan bahasa adalah lebih memuaskan (.40) apabila kaedah holistik digunakan untuk menilai sampel karangan tersebut. Ini menunjukkan bahawa kesan tugasan penulisan dan pemeriksa banyak bersandar kepada kaedah dan aspek pemarkahan yang dipilih.

Setelah membuat rujukan ke atas kajian-kajian lepas yang dijalankan oleh para penyelidik sebelumnya mengenai faktor tugasan, faktor pemeriksa dan faktor pemarkahan, satu kerangka kajian telah dibina (Rajah 2.1) untuk menjelaskan interaksi faktor-faktor tersebut dan pengaruhnya terhadap skor karangan.



Rajah 2.1. Interaksi sumber variasi dan pengaruhnya terhadap kebolehpercayaan skor karangan.

2.2 Sorotan Kajian Lepas Tentang Pentaksiran Karangan Yang Menggunakan Teori G

Chen, Niemi, Wang, Wang, dan Mirocha (2007) telah menjalankan satu kajian tentang kebolehan generalisasi tugasan penulisan terhadap 397 orang pelajar Gred 9 daripada 19 kelas di satu sekolah bandar. Tujuan kajian mereka adalah untuk meneroka tahap kebolehan generalisasi beberapa tugasan yang dianggap berkualiti tinggi dan memantapkan suatu kesahan yang boleh digunakan untuk mentaksir

kebolehan menulis pelajar dengan menggunakan bilangan tugasan karangan yang terhad.

Dalam penelitian semula kajian-kajian lepas yang dijalankan, para penyelidik merumuskan bahawa prestasi penulisan pelajar sangat bergantung kepada sejauh mana kebiasaannya pelajar (mempunyai pengetahuan yang mencukupi) terhadap sesuatu topik. Oleh itu, mereka menjalankan kajian ini untuk melihat pencapaian prestasi penulisan pelajar menerusi pengawalan rangsangan karangan yang berbeza iaitu memberi rangsangan karangan berunsur kebiasaan bagi pelajar, dan bilangan karangan yang patut digunakan untuk mendapat keputusan yang boleh dipercayai tentang kebolehan menulis pelajar.

Dalam kajian mereka, empat tugasan karangan yang berbeza (dua daripadanya disediakan petikan bacaan untuk mengawal asas pengetahuan pada takat tertentu) telah digunakan. Tiga daripada tugasan tersebut adalah bahan berasaskan karya sastera yang pelajar telah baca dalam bilik darjah dan satu lagi ialah cerpen bukan fiksyen. Semua rangsangan karangan disediakan oleh CRESST (*National Centre for Research of Evaluation, Standards & Student Testing, University of California*). Setiap pelajar perlu menulis dua buah karangan. Karangan yang pertama wajib dijawab. Karangan kedua dipilih daripada mana-mana tugasan karangan yang diberikan selebihnya. Kedua-dua karangan perlu disiapkan dalam waktu kelas Bahasa English dengan setiap satu karangan diberi peruntukan masa menulis selama dua waktu kelas. Selepas penulisan karangan pertama, karangan kedua perlu disiapkan dalam tempoh masa dua minggu untuk mengelakkan kesan kematangan.

Untuk mengimbang balas kesan susunan tugasan, setiap pelajar daripada kelas yang berlainan akan diberi rawatan keadaan eksperimen secara rawak.

Empat orang bekas guru Bahasa English sekolah tinggi telah diberi sesi latihan menanda karangan selama tiga jam. Rubrik penskoran holistik (skor 1 hingga 4) yang dibentuk oleh CRESST telah digunakan oleh pemeriksa untuk menilai karangan mengikut aspek karangan iaitu kandungan, organisasi, kefahaman dan mekanis. Kebolehpercayaan pemeriksa telah disemak semasa dan selepas latihan pemeriksaan. Setiap karangan diberi skor oleh empat pemeriksa dan purata skor karangan digunakan dalam analisis data. Untuk mendapatkan keputusan pekali G, reka bentuk kajian G iaitu reka bentuk dua faset tersilang $p \times t \times r$ kesan rawak telah dijalankan di mana p ialah pelajar manakala t dan r masing-masing adalah tugasan dan pemeriksa.

Keputusan karangan direkod mengikut latar belakang jantina, etnik, pelajar berbahasa English, pendidikan khas dan kumpulan bestari. Hasil kajian menunjukkan skor min untuk pelajar perempuan lebih tinggi daripada pelajar lelaki, kumpulan bestari lebih tinggi berbanding dengan kumpulan lain. Etnik Spanish mendapat min lebih rendah daripada etnik berkulit putih dan etnik lain. Pelajar berbahasa English dan pelajar pendidikan khas mendapat min yang lebih rendah berbanding dengan pelajar lain. Korelasi skor antara pelajar bagi karangan pertama dan skor bagi karangan lain mempunyai nilai yang sederhana iaitu antara .61 hingga .68.

Tiga kajian G dengan reka bentuk yang sama telah dijalankan bagi membandingkan skor karangan 1 dan karangan 2, karangan 1 dan karangan 3 serta

karangan 1 dan karangan 4 untuk golongan pelajar yang sama. Analisis komponen varian telah dijalankan bagi setiap kajian untuk menguji sumber varian dalam karangan. Keputusan kajian G yang diperoleh adalah hampir sama. Berdasarkan cerapan tunggal, pekali G bagi tiga kajian ini adalah seperti berikut: Kajian 1 (.37), Kajian 2 (.38) dan Kajian 3 (.40). Secara umum, varian skor karangan bagi perbezaan kebolehan menulis σ^2 (*p*) adalah antara 36% hingga 40%. Sumber varian signifikan yang seterusnya adalah interaksi pelajar × tugasan × pemeriksa, σ^2 (*ptr*) (34% hingga 40%). Varian daripada interaksi pelajar × tugasan σ^2 (*pt*) adalah antara 11% hingga 18% daripada jumlah varian. Ini menunjukkan pencapaian pelajar adalah berbeza mengikut tugasan karangan yang berlainan. Komponen varian σ^2 (*tr*) adalah hampir kosong. Ini menunjukkan pemarkahan yang dilakukan oleh empat orang pemeriksa adalah boleh dipercayai. Dalam Kajian 2 dan Kajian 3, komponen varian bagi σ^2 (*pt*) adalah lebih besar (13.9% dan 18.2%) daripada komponen varian bagi σ^2 (*pr*) (4.7% dan 6.4%). Namun begitu, dalam Kajian 1 pula, komponen varian bagi σ^2 (*pr*) adalah lebih besar (16.9%) daripada σ^2 (*pt*) (10.7%).

Kajian D dalam kajian ini menunjukkan gabungan dua buah karangan yang berlainan tugasan dan empat orang pemeriksa adalah tidak mencukupi untuk menilai kebolehan menulis pelajar. Untuk mencapai pekali G pada tahap .80, Kajian 2 dan Kajian 3 memerlukan gabungan tiga buah karangan dan empat orang pemeriksa ($\acute{n}_t = 3$, $\acute{n}_r = 4$) manakala Kajian 1 memerlukan lima buah karangan dan empat orang pemeriksa ($\acute{n}_t = 5$, $\acute{n}_r = 4$) kerana komponen varian σ^2 (*pr*) adalah lebih besar dalam kajian ini. Secara kesimpulan kajian ini menunjukkan ralat pengukuran dalam tugasan adalah lebih besar daripada pemeriksa, maka pertambahan bilangan tugasan adalah lebih berkesan daripada penambahan pemeriksa dalam meningkatkan

kebolehpercayaan skor.

Chen dan rakan-rakannya (2007) mendakwa terdapat dua kemungkinan berlakunya *generalizability* yang rendah dalam kajian ini. Pertama, kelemahan rubrik penskoran holistik yang hanya memberikan skor tunggal berbanding dengan pemarkahan analitik yang menggunakan pelbagai ciri untuk mentaksir prestasi penulisan pelajar mungkin lebih berupaya meningkatkan *generalizability* tugasan. Selain itu, julat skala pemarkahan yang sempit iaitu satu hingga empat poin dan penumpuan pemarkahan pada dua dan tiga poin yang diberikan oleh pemeriksa dalam kajian ini. Oleh itu, mereka mencadangkan penggunaan julat skala yang lebih panjang dengan arahan kepada pemerika berserta dengan penggunaan rubrik pemarkahan analitik untuk meningkatkan *generalizability* tugasan. Namun begitu, tugasan yang digubal dengan terperinci dalam kajian ini tidak semestinya menjamin kebolehpercayaan skor yang tinggi. Faktor keberubahan tentang prestasi pelajar terhadap tugasan karangan yang berlainan walaupun bagi tugasan daripada domain yang sama perlu juga dipertimbangkan (Brennan, 1996).

Satu lagi kajian tentang penggunaan teori G telah dilakukan oleh Brown (2007). Beliau telah menggunakan teori G untuk mengkaji pengaruh bilangan jenis tajuk dan bilangan pemarkahan (*ratings*) terhadap kebergantungan skor karangan dalam bahasa pertama berdasarkan skor karangan daripada *Manoa Writing Placement Examination* (MWPE). Untuk tujuan tersebut, beliau menganalisis skor 6875 orang pelajar ijazah muda yang telah mengambil peperiksaan MWPE dalam jangka masa empat tahun yang lalu.

Dalam peperiksaan MWPE yang bercirikan rujukan norma, setiap pelajar dikehendaki menulis dua buah karangan yang berlainan tajuk. Karangan pertama dikehendaki pelajar membaca sebuah artikel sepanjang dua muka surat kemudian menuliskan respons mereka. Karangan kedua adalah berkaitan dengan pengalaman diri yang berunsur naratif. Sebanyak 24 set tajuk karangan yang digubal dalam jangka masa empat tahun yang lalu telah digunakan sebagai bahan ujian. Setiap pelajar diagihkan secara rawak dengan satu set tajuk karangan tertentu. Untuk tujuan imbang balas (*counterbalance*), separuh daripada pelajar akan menjawab tajuk membaca dan menulis dan separuh lagi menjawab tajuk pengalaman diri. Mereka diberi masa 270 minit untuk menulis dua buah karangan tersebut iaitu 75 minit untuk mendraf karangan dan 60 minit untuk menyemak bagi setiap karangan.

Dalam kajian ini, karangan pelajar dinilai berdasarkan enam aspek iaitu tesis / bukti, organisasi / perkembangan, tatabahasa, stail, ejaan dan tanda baca dengan pemarkahan holistik yang menggunakan bacaan cepat. Skala pemarkahan 0 hingga 5 poin digunakan. Setiap satu tajuk karangan dimarkah dua kali oleh dua orang pemeriksa yang berlainan. Pemeriksa adalah ahli fakulti yang mempunyai pengalaman mengajar di Universiti Hawaii. Sebelum pemarkahan skrip, pemeriksa telah diberi satu bungkusan bahan penskoran untuk membiasakan mereka dengan prosedur pemarkahan. Pemeriksa diberi masa tiga jam untuk menyiapkan tugas pemeriksaan termasuk satu jam untuk latihan pemarkahan. Jika perbezaan skor pemarkahan adalah lebih daripada dua poin, maka pemeriksa ketiga akan menjadi pengadil. Dalam kajian ini, perbezaan seperti ini adalah kira-kira 6%.

Reka bentuk $p \times (r:t)$ telah digunakan dalam kajian G di mana *p*, *r* dan *t* masing-masing mewakili pelajar, pemarkahan dan jenis tajuk karangan. Kedua-dua faset tajuk dan pemarkahan dalam kajian tersebut dirawat sebagai faset rawak kerana jenis tajuk dan pemeriksa adalah diandaikan sebagai sampel-sampel rawak daripada semua kemungkinan jenis tajuk dan pemeriksa dalam semesta masing-masing (lihat Brown & Hudson, 2002, p.189) dan ini telah dijustifikasikan berdasarkan konsep saling bertukaran (*exchangeability*) yang dibincangkan oleh Shavelson dan Webb (1991, p.11-12). Keputusan kajian G menerusi prosedur ANOVA untuk setiap komponen varian adalah seperti dalam Jadual 2.2. Komponen varian $\sigma^2(prt)$ adalah paling besar dan sumbangan faset pemeriksa kepada jumlah varian adalah sangat kecil. Komponen varian $\sigma^2(pt)$ yang agak besar (.16) menunjukkan bahawa kedudukan relatif bagi skor pelajar adalah berlainan bagi tugasan yang berbeza. Pekali G dalam kajian ini adalah kira-kira .296 berdasarkan cerapan tunggal. Manakala berdasarkan reka bentuk asalnya iaitu $t = 2$ dan $r = 4$, pekali G adalah kira-kira .632. Keputusan kajian D menunjukkan bilangan tajuk mempunyai kesan pengaruh yang lebih besar ke atas kebolehpercayaan skor daripada bilangan pemarkahan.

Jadual 2.2

Anggaran Komponen-Komponen Varian Bagi Reka Bentuk p ×(r: t) Kajian G

Sumber	*df*	*SS*	*MS*	σ^2
Person (*p*)	6874	10495.8097	1.5269	0.2139647
Tajuk (*t*)	1	0.8624	0.8624	0.0000000
Pemarkahan:Tajuk (*r:t*)	2	2.5543	1.2772	0.0001353
Person × *Tajuk* (*pt*)	6874	4612.6376	0.6710	0.1620537
Person × *Pemarkahan:Tajuk* (*prt,e*)	13748	4769.4457	0.3469	0.3469192

Sumber: Daripada Brown, J. D. (2007, p. 14). Multiple views of L1 writing score reliability. *Second Language Studies, 25(2)*, 1-31.
Nota. σ^2 = Anggaran Komponen Varian.

Manakala dalam kajian yang lain, Schoonen (2005) telah menjalankan kajian *generalizability* tentang skor penulisan dengan menggunakan model SEM (*Structural Equation Modeling*). Tujuannya adalah untuk mengkaji kesan pemeriksa dan tugasan karangan terhadap generalizalibiti skor penulisan. Beliau juga menganalisis perkaitan kesan tersebut melalui dua aspek penilaian dan dua kaedah penskoran yang digunakan dalam kajiannya. Selain daripada itu, kajian ini juga menunjukkan penggunaan model SEM dalam menganggarkan komponen varian yang boleh diperoleh dalam konteks kajian G.

Kajian ini telah dijalankan di 22 buah sekolah di Belanda. Subjek kajian terdiri daripada pelajar Gred 6 yang berumur antara 11 hingga 12 tahun pada tahun akhir peringkat sekolah rendah. Seramai 89 orang pelajar telah dipilih secara rawak daripada 442, iaitu jumlah keseluruhan pelajar Gred 6 yang terdapat di sekolah-sekolah tersebut. Subjek kajian telah diminta menyiapkan empat buah tugasan karangan iaitu sebuah karangan pemujukan, sebuah karangan deskriptif dan dua buah karangan arahan atau pendedahan. Tugasan penulisan tersebut asalnya dibentuk bagi kegunaan pentaksiran peringkat kebangsaan. Situasi komunikasi yang spesifik bagi tugasan tersebut iaitu motif penulisan, matlamat yang perlu dicapai dan sasaran pembaca yang dikehendaki telah dijelaskan secara terperinci kepada subjek kajian. Di samping itu, subjek juga diberikan bahan rangsangan tertentu seperti artikel surat khabar, lukisan dan gambar rajah untuk tujuan pemahaman tentang kandungan penulisan.

Semua tugasan ini diberi skor mengikut dua aspek penilaian, iaitu aspek pertama melibatkan kandungan dan organisasi, manakala aspek kedua adalah

penggunaan bahasa. Setiap aspek diberi skor mengikut dua kaedah penskoran yang berlainan iaitu penskoran secara holistik dan penskoran secara analitik. Penskoran holistik menggunakan rujukan *benchmark* karangan berskala lima manakala penskoran analitik menggunakan panduan penskoran yang ketat untuk menentukan sama ada hadir atau tidak sesuatu proposisi yang berkaitan (untuk kandungan dan organisasi) atau kesilapan bahasa (untuk penggunaan bahasa). Penskoran tugasan karangan telah dilakukan oleh enam panel yang setiap satunya mengandungi lima orang pemeriksa. Setiap panel tidak akan mengulang prosedur pemarkahan yang sama. Tugasan pemarkahan yang berbeza dijalankan pada sesi pemarkahan yang berlainan. Pemeriksa telah dipilih secara rawak dan mereka mempunyai latar belakang pendidikan atau profesional dalam linguistik gunaan ataupun sebagai guru. Kebanyakan pemeriksa juga mempunyai pengalaman profesional sebagai guru sekolah rendah.

Berdasarkan reka bentuk kajian ini, setiap pelajar memperolehi 80 skor untuk penulisan yang dibuat iaitu 4 tugasan × 2 aspek penilaian × 2 kaedah penskoran × 5 pemeriksa. Min untuk penskoran holistik ditetapkan pada 100 dan sisihan piawai pada 15. Keempat-empat tugasan penulisan secara puratanya sama ada berdasarkan isi kandungan dan organisasi atau penggunaan bahasa mempunyai min antara 93.47 hingga 111.61, dan sisihan piawai adalah antara 11.50 hingga 16.92. Manakala dalam penskoran analitik, minnya adalah antara 3.04 hingga 7.99 dan sisihan piawai antara 1.12 hingga 3.24. Memandangkan kedua-dua cara penskoran tersebut menggunakan skala yang agak berlainan, Schoonen (2005) telah menganalisis prosedur penskoran secara berasingan dengan menggunakan SEM. Menurut beliau, tugasan karangan dan prosedur penskoran digunakan dalam pentaksiran rujukan

norma. Ini bermakna kajiannya adalah tertumpu kepada ralat relatif iaitu melihat interaksi antara pelajar dengan faset tugasan karangan. Ralat relatif hanya mempengaruhi kedudukan pelajar atau objek pengukuran.

Dalam kajian ini, Schoonen (2005) menggunakan reka bentuk separa tersarang $p \times (r: t)$ kesan rawak iaitu pemeriksa tersarang dalam tugasan yang berlainan dan semua pelajar menjawab tugasan yang diberikan sementara tugasan dan pemeriksa adalah tersilang dengan pelajar dalam kajian G dan kajian D untuk membuat keputusan relatif. Pada keseluruhannya, berdasarkan cerapan tunggal, pekali G bagi aspek penggunaan bahasa (kaedah holistik 0.40, kaedah analitik 0.30) secara relatif adalah lebih tinggi berbanding dengan aspek kandungan dan organisasi (kaedah holistik 0.32, kaedah analitik 0.21). Manakala pekali G bagi kaedah holistik adalah lebih tinggi berbanding dengan kaedah analitik. Pekali yang rendah ini menunjukkan terdapat korelasi yang rendah antara skor penulisan.

Kajian G menunjukkan dari segi penskoran kandungan dan organisasi, peratus varian σ^2 (*pt*) adalah terbesar sekali (kaedah holistik, 41.4%; kaedah analitik, 53.0%) daripada jumlah varian berbanding dengan kesan pelajar σ^2 (*p*) (kaedah holistik, 32.0%; kaedah analitik, 20.5%). Manakala varian reja σ^2 (*prt,e*) yang juga mengandungi varian σ^2 (*pr*) dan σ^2 (*prt*) adalah secara relatif kecil sedikit. Pola ini lebih jelas dalam penskoran analitik daripada penskoran holistik. Schoonen (2005) mendakwa keadaan ini adalah disebabkan pemarkahan holistik yang memirip kepada ciri am kurang bergantung kepada tugasan dalam penskoran kandungan dan organisasi. Kesan σ^2 (*pt*) yang besar menandakan kedudukan pelajar sangat bergantung kepada tugasan karangan. Berbanding dengan penskoran aspek

kandungan dan organisasi, keputusan penskoran penggunaan bahasa adalah lebih baik sedikit memandangkan komponen varian σ^2 (*p*) secara relatif adalah lebih besar (kaedah holistik, 39.6%; kaedah analitik, 29.4%). Manakala varian σ^2 (*pt*) adalah lebih kecil dalam pemarkahan holistik (20.6%) berbanding dengan pemarkahan analitik (33.1%). Kesan σ^2 (*pt*) yang agak besar ini menunjukkan pengiraan kesilapan bahasa juga dipengaruhi oleh tugasan karangan.

Kajian ini mendapati aspek penggunaan bahasa lebih mudah mencapai pekali G .80 yang dihasratkan berbanding dengan aspek kandungan dan organisasi. Pekali G .80 juga lebih senang dicapai oleh kaedah penskoran holistik berbanding dengan kaedah penskoran analitik. Dalam kaedah penskoran analitik untuk aspek kandungan dan organisasi, walaupun 10 tugasan dan 10 pemeriksa digunakan, pekali G yang dicapai masih pada .787. Manakala dalam penskoran holistik untuk penggunaan bahasa, pekali G .80 dicapai dengan tiga tugasan disemak oleh lima pemeriksa ($E\rho^2$=.806) ataupun empat tugasan disemak oleh tiga pemeriksa ($E\rho^2$=.824). Berdasarkan reka bentuk kajian asal iaitu empat tugasan dan lima pemeriksa digunakan, pekali G .80 hanya dapat dicecah dalam prosedur penskoran holistik untuk penggunaan bahasa ($E\rho^2$=.847) manakala penskoran analitik untuk aspek kandungan dan organisasi mencapai pekali G .585 sahaja. Kajian ini juga menunjukkan pertambahan bilangan tugasan akan membawa kepada peningkatan pekali G yang lebih besar walaupun sehingga lima atau enam tugasan digunakan terutamanya dalam pemarkahan kandungan dan organisasi. Walau bagaimanapun, penggunaan bilangan pemeriksa yang lebih daripada dua atau tiga orang tidak banyak membantu dalam peningkatan pekali G.

Dapatan kajian Schoonen (2005) membuktikan sebahagian besar skor karangan dipengaruhi oleh kesan faset pentaksiran iaitu tugasan karangan dan pemeriksa selain daripada kebolehan menulis pelajar. Namun begitu, kesan pengaruh ini bergantung kepada aspek dan kaedah penskoran. Selain itu, kebolehan generalisasi bagi skor analitik adalah lebih lemah berbanding dengan skor holistik manakala kebolehan generalisasi bagi skor pengguanan bahasa adalah lebih kuat daripada skor kandungan dan organisasi. Dalam kajian ini, subjek dikehendaki menulis empat buah karangan dalam tiga bentuk karangan yang berlainan iaitu pemujukan, deskriptif dan arahan atau pendedahan. Namun begitu, tidak ada bukti dari segi data menunjukkan bahawa dua buah karangan yang berbentuk pendedahan tersebut mempunyai min yang hampir sama antara satu sama lain berbanding dengan karangan bentuk pemujukan dan deskriptif. Walaupun Schoonen (2005) mengakui bahawa ia adalah satu ujian yang lemah untuk menolak hipotesis tentang kesan bentuk karangan mempengaruhi prestasi penulisan, namun kenyataan tersebut mempunyai konotasi tertentu iaitu untuk mengukuhkan dakwaannya bahawa varian tugasan spesifik (*task-specific*) adalah varian ralat dalam kajiannya dan bukannya varian yang disebabkan oleh bentuk karangan.

Lee dan Kantor (2005) telah menjalankan kajian untuk menilai kebergantungan skor bagi pentaksiran penulisan ESL (*English as a second language*) format baru dengan menggunakan dua jenis tugasan prototaip berdasarkan reka bentuk skema penilaian yang berbeza. Tujuan kajian adalah untuk meneliti kesan relatif tugasan dan pemeriksa ke atas skor penulisan berpandukan tugasan gabungan dan tugasan tunggal di samping melihat impak gabungan tentang bilangan tugasan dan pemeriksa, dan juga impak tentang bilangan penilaian iaitu penilaian tunggal dan

dua kali terhadap kebergantungan skor dari perspektif teori G. Tugasan-tugasan prototaip yang digunakan dalam kajian ini terdiri daripada dua jenis tugasan gabungan iaitu mendengar-menulis (LW) dan membaca-menulis (RW), dan satu tugasan tunggal (IW). Tugasan gabungan mentaksir kebolehan calon menghasilkan esei dengan mengintegrasikan pelbagai aspek kemahiran bahasa mereka melalui bahan rangsangan seperti syarahan atau teks akademik. Manakala tugasan tunggal yang diuji bertujuan mencungkil pengalaman peribadi atau pengetahuan am calon semasa memberi respons kepada rangsangan penulisan. Sebanyak lapan jenis tugasan penulisan iaitu tiga LW, dua RW dan tiga IW telah ditadbirkan dalam kajian ini. Berbeza sedikit dengan kajian-kajian lain, kajian ini cuba meneroka skema penilaian karangan yang baru iaitu semua karangan dinilai sekali sahaja, tetapi setiap tugasan bagi calon akan dinilai oleh pemeriksa yang berlainan. Skema penilaian tunggal ini menyebabkan bilangan pemeriksa berkurangan daripada dua kepada satu, tetapi bilangan pemeriksa per calon adalah sama seperti bilangan tugasan yang diberikan dalam ujian. Ini kerana kebanyakan pentaksiran berasaskan prestasi yang berskala besar menggunakan dua orang pemeriksa terlatih untuk menilai satu atau lebih tugasan yang dihasilkan oleh setiap calon.

Kajian ini terbahagi kepada dua fasa. Dalam fasa pertama, seramai 488 subjek kajian (233 lelaki, 247 perempuan dan 8 tidak dikenal pasti jantina) yang berumur secara purata 22.3 tahun terlibat dalam kajian ini. Mereka adalah pelajar ESL/EFL (*English as a second / foreign language*) yang dipilih daripada tiga pusat peperiksaan dalam Amerika dan lima dari luar negara iaitu Australia, Kanada, Hong Kong, Mexico dan Taiwan. Mereka juga telah menduduki kertas TOEFL versi ITP (*Institutional Testing Program*) dan memperoleh skor dalam lingkungan 337 hingga

673 (min = 558; sisihan piawai = 61). Latar belakang bahasa subjek kajian adalah pelbagai termasuklah lima kumpulan bahasa ibunda terbesar iaitu Cina (26%), Sepanyol (22%), Kantonis (11%), Korea (8%), dan Thai (7%). Subjek kajian dibahagi kepada tiga subkumpulan (n_p= 162 untuk subkumpulan 1, n_p = 164 untuk subkumpulan 2, n_p = 162 untuk subkumpulan 3). Semua subkumpulan perlu menjawab enam tugasan penulisan iaitu tiga tugasan LW dan dua tugasan RW selain satu tugasan IW yang dikhaskan kepada setiap subkumpulan. Dalam fasa kajian ini, setiap respons calon dinilai sebanyak dua kali berdasarkan pendekatan holistik yang berskala 1 hingga 5. Pasangan pemeriksa yang berbeza dipilih secara rawak daripada sekumpulan 27 pemeriksa dan mereka diagihkan tugas untuk menilai setiap esei dalam tugasan tersebut.

Dalam kajian fasa pertama, analisis *univariate* secara berasingan dijalankan untuk menganggar komponen varian untuk kesan utama dan kesan interaksi bagi tiga subkumpulan berdasarkan set data asal. Setiap komponen varian yang sama merentas tiga subkumpulan telah dipuratakan supaya mendapatkan anggaran yang lebih tekal (Brennan et al., 1995; Gao, Shavelson, & Baxter, 1994), dan semua komponen varian yang dipuratakan tersebut digunakan untuk menganggar pekali G dan pekali D bagi bahagian-bahagian penulisan dalam kajian ini. Dua jenis reka bentuk kajian G telah digunakan untuk menganggar komponen varian untuk kesan utama dan kesan interaksi. Pertama, reka bentuk dua faset tersilang ($p \times t \times r'$) dengan tugasan (t) dan bilangan penilaian (r') sebagai faset rawak (n_p= 488, n_t= 6, $n_{r'}$ = 2) dan kedua, reka bentuk dua faset separa tersarang [(r:p) × t] dengan tugasan (t) dan pemeriksa (r) sebagai faset rawak (n_p= 488, n_t= 6, n_r = 2). Dapatan kajian G menunjukkan bahawa nilai komponen varian yang utama bagi kedua-dua reka bentuk tersebut adalah

hampir-hampir sama : komponen varian σ^2 (*p*) (.51 atau 39.1%), σ^2 (*ptr'*) (.38 atau 29.5%) atau σ^2 (*tr*:*p*) (.38 atau 29.6%) , σ^2 (*pt*) (.26 atau 20.2%) dan komponen varian σ^2 (*t*) (.14 atau 10.6%). Manakala nilai komponen varian yang berkaitan dengan kesan pemeriksa (pemarkahan) kecuali interaksi kesan tiga hala adalah kecil sahaja. Memandangkan kedua-dua reka bentuk kajian G menghasilkan keputusan yang hampir sama, Lee dan Kantor (2005) hanya melaporkan analisis kajian D bagi reka bentuk dua faset tersilang ($p \times t \times r'$). Berdasarkan bilangan tugasan 1 hingga 10, reka bentuk penilaian tersebut merekodkan pekali G antara .44 hingga .88 bagi penilaian tunggal per esei dan pekali G bagi penilaian dua kali per esei adalah antara .53 hingga .91. Manakala berdasarkan data asal (n_p= 488, n_t= 6, $n_{r'}$ = 2), pekali G adalah .82 dan .87 bagi penilaian tunggal dan dua kali per esei masing-masing. Selain itu, laporan pekali kebolehpercayaan bagi reka bentuk satu faset tersilang ($p \times t$) dengan bilangan penilaian (*r'*) sebagai faset tersembunyi berdasarkan bilangan tugasan 1 hingga 10 juga dikemukakan dan didapati nilai pekali yang diperoleh (antara .53 hingga .92) adalah sangat hampir dengan nilai pekali bagi senario penilaian dua kali per esei.

Oleh kerana data asal bagi reka bentuk penilaian $p \times t \times r'$ dan $(r:p) \times t$ dalam fasa pertama tidak membenarkan kesan utama *r* dan interaksi *pr* dianggar secara berasingan dan kedua-dua reka bentuk tersebut juga bukanlah perwakilan tipikal untuk keadaan penilaian atas talian dikendalikan ETS, Lee dan Kantor telah menjalankan kajian fasa kedua. Dalam fasa ini, reka bentuk tersilang ($p \times t \times r$) yang mirip dengan keadaan penilaian atas talian ETS bagi pentaksiran penulisan skala besar telah digunakan. Setiap esei bagi calon subkumpulan 3 (n_p = 162) telah dinilai semula oleh 6 orang pemeriksa terlatih yang dipilih daripada sekumpulan 27 orang

pemeriksa yang terlibat dalam sesi pemarkahan dalam kajian fasa pertama. Mereka dipilih kerana mempunyai kebolehpercayaan antara pemeriksa yang tinggi. Untuk mengurangkan kesan halo, mereka diminta menilai jenis tugasan yang sama bagi semua calon sebelum meneruskan pemarkahan tugasan yang lain. Masa untuk menyiapkan tugas pemarkahan adalah seminggu. Dapatan kajian G menunjukkan amaun varian berkaitan dengan tugasan dan pemeriksa (bilangan penilaian) bagi data dinilai semula ($p \times t \times r$, n_p= 162, n_t = 6, n_r = 6) dan data asal ($p \times t \times r'$, n_p= 162, n_t = 6, $n_{r'}$ = 2) memaparkan pola yang sama iaitu komponen varian σ^2 (p) adalah paling besar, diikuti oleh σ^2 (ptr) atau σ^2 (ptr'), dan σ^2 (pt). Dalam kajian ini, dua senario penilaian yang diberi perhatian dalam kajian D ialah tiga tugasan dinilai sekali berdasarkan reka bentuk $R:$ ($p \times T$) dan satu tugasan dinilai dua kali berdasarkan ($R:p$) $\times$ T dengan masing-masing memperoleh nilai pekali G sebanyak .77 dan .62. Berdasarkan bilangan tugasan 1 hingga 10 berdasarkan penilaian tunggal per esei, reka bentuk penilaian $R:$ ($p \times T$) dan ($R:p$) $\times$ T memperoleh pekali G antara .53 hingga .92 dan antara .53 hingga .88 masing-masing manakala bagi bilangan tugasan yang sama dengan penilaian dua kali per esei, pekali G yang dicapai adalah antara .59 hingga .94 dan antara .62 hingga .92. Selain itu, kebolehpercayaan skor bagi reka bentuk penilaian $p \times T \times R$ dan $p \times (R: T)$ berdasarkan data dinilai semula, dan reka bentuk penilaian $p \times T \times R'$ berdasarkan data asal untuk sampel kajian keseluruhan dan subkumpulan 3 juga dijadikan sebagai asas perbandingan.

Beberapa dapatan kajian dapat disimpulkan berdasarkan kajian ini. Kajian ini mendapati bahawa sumber varian yang paling besar bagi prestasi calon adalah berpunca daripada perbezaan kebolehan calon atau σ^2 (p). Selain itu, cara yang lebih berkesan untuk memaksimumkan kebergantungan skor adalah untuk menambahkan

bilangan tugasan daripada bilangan penilaian per esei atau pemeriksa. Dalam kajian ini, antara empat reka bentuk penilaian yang dikemukakan berdasarkan senario penilaian tunggal per esei [$p \times T \times R$, (R:p) $\times$ T, $p \times (R{:}T)$, R: ($p \times T$)] dalam kajian D, didapati reka bentuk penilaian yang menggunakan tugasan berbeza untuk calon yang sama kemudian dinilai oleh pemeriksa yang berbeza [$p \times (R{:}\ T)$, R: ($p \times T$)] memperoleh kebergantungan skor yang lebih tinggi bagi keadaan kombinasi jenis penulisan yang berbeza terutamanya dalam keadaan dua atau lebih tugasan digunakan. Namun begitu, kelebihan kedua-dua reka bentuk tersebut terhapus apabila penilaian dua kali per esei digunakan. Kajian ini juga mendapati kebolehpercayaan skor adalah lebih tinggi bagi kombinasi jenis penulisan apabila lebih banyak bilangan tugasan LW digunakan berbanding dengan tugasan RW. Kajian ini juga memaparkan aspek generalisasi bagi tugasan mungkin merupakan cabaran yang dihadapi oleh pentaksiran penulisan TOEFL format baru. Ini kerana tugasan yang diuji adalah berbeza dari segi kesukaran dan tugasan tersebut juga menunjukkan tahap kesukaran yang berlainan bagi semua calon. Namun begitu, varian bagi faset pemeriksa secara relatif adalah kecil sahaja. Ketegasan pemarkahan di kalangan pemeriksa adalah tidak banyak berbeza, begitu juga ketegasan pemeriksa terhadap calon pada keseluruhannya.

Sementara itu, Sudweeks, Reeve dan Bradshaw (2004) telah menjalankan satu kajian rintis untuk menilai dan membaiki prosedur pemarkahan karangan untuk mentaksir kebolehan menulis bagi pelajar kolej Tahun Dua berdasarkan teori G dan *many-facet Rasch measurement* (MFRM). Sampel karangannya didapati daripada 497 orang pelajar ijazah muda kursus sejarah dunia. Setiap pelajar dikehendaki menulis dua buah karangan atas tajuk tertentu yang relevan dengan sejarah dunia dan

panjangnya tiga muka surat. Sebanyak 48 sampel karangan yang dihasilkan oleh 24 pelajar telah dipilih sebagai bahan kajian. Sampel karangan tersebut juga mewakili pelbagai tahap kebolehan menulis pelajar. Sembilan orang pemeriksa daripada Jabatan Bahasa Inggeris diminta menilai 48 sampel karangan tersebut sebanyak dua kali berdasarkan dua keadaan (*occasion*) pemarkahan yang berbeza. Pemeriksa-pemeriksa tersebut menilai kedua-dua buah karangan yang dihasilkan oleh setiap pelajar. Pendekatan pemarkahan holistik yang berskala sembilan poin digunakan untuk menilai karangan tersebut. Setiap poin mempunyai tajuk deskriptif tertentu dan disertakan dengan satu atau dua ayat untuk menerangkan ciri karangan pada poin tersebut. Para pemeriksa diberi latihan pemarkahan sebanyak dua jam dalam kedua-dua sesi latihan. Selepas itu, mereka menilai beberapa buah karangan yang berlainan mutu dengan maklum balas daripada kumpulan ahli sebagai latihan pengukuhan setelah memahami kandungan rubrik pemarkahan.

Kajian G dijalankan dengan menggunakan reka bentuk tiga faset tersilang sepenuhnya iaitu setiap pemeriksa akan menilai semua tugasan karangan pelajar berdasarkan bilangan keadaan pemarkahan yang berlainan. Pada keseluruhannya terdapat 15 sumber keberubahan dalam reka bentuk tersebut. Output daripada analisis kajian G menunjukkan bahawa komponen varian $\sigma^2(p)$ adalah paling besar (28.8%) manakala komponen varian yang berkaitan dengan tugasan penulisan merangkumi hampir 40% daripada jumlah varian [$\sigma^2(t)$ = 9.8%, $\sigma^2(pt)$ = 14.9%, $\sigma^2(ptr)$ = 14.4%]. Selain itu, kesan reja atau baki varian yang tidak dapat dijelaskan merupakan 22.3% daripada jumlah varian dan adalah komponen varian kedua terbesar. Komponen varian $\sigma^2(pt)$ yang agak besar dalam kajian ini menunjukkan sebarang generalisasi tentang kedudukan relatif pelajar berdasarkan salah satu

daripada tugasan karangan adalah kurang boleh dipercayai dan mungkin memberi kesimpulan yang berlainan terhadap kebolehan menulis pelajar. Ini dikukuhkan dengan keadaan interaksi tiga hala $\sigma^2(ptr)$ yang agak besar yang menandakan interaksi dua hala *pt* adalah tidak sama merentas pemeriksa yang berlainan. Justeru itu, penggunaan tugasan yang lebih banyak adalah perlu agar generalisasi yang boleh dipercayai tentang kebolehan menulis pelajar dapat dilakukan. Selain itu, para penyelidik menyatakan kebimbangan tentang kesan reja yang besar walaupun mereka telah mengambil kira pelbagai sumber ralat yang terlibat dalam prosedur pemarkahan.

Untuk menganggar kebolehpercayaan tentang pemarkahan yang diperoleh, dan mengunjur kemungkinan perubahan kebolehpercayaan dengan menggunakan bilangan pemeriksa, tugasan dan keadaan pemarkahan yang berlainan, empat jenis reka bentuk kajian yang berlainan [$p\times T\times R$, $p \times (R{:}T)$, $(R{:}p) \times T$ dan $R{:}(p \times T)$] telah digunakan dalam analisis kajian D bagi kajian ini. Kajian ini telah melaporkan nilai pekali G bagi bilangan tugasan dan pemeriksa satu hingga empat serta dua keadaan pemarkahan berdasarkan reka bentuk kajian yang berlainan. Untuk reka bentuk tersilang $p\times T\times R$ yang mengandaikan setiap pemeriksa akan menilai kedua-dua tugasan bagi semua pelajar, masing-masing mencatatkan nilai pekali G dalam lingkungan .34 hingga .79 (keadaan pemarkahan tunggal) dan .39 hingga .81 (dua keadaan pemarkahan). Untuk reka bentuk kajian separa tersarang $p \times (R{:}T)$ pula iaitu setiap pemeriksa akan menilai tugasan tertentu bagi semua pelajar, nilai pekali G bagi keadaan pemarkahan tunggal adalah antara .43 hingga .83 manakala berdasarkan dua keadaan pemarkahan adalah antara .52 hingga .86. Untuk reka bentuk kajian separa tersarang $(R{:}p) \times T$ iaitu setiap pemeriksa akan menilai kedua-

dua tugasan bagi pelajar tertentu, dapatan pekali G bagi senario pemarkahan tunggal adalah antara .32 hingga .78 dan antara .37 hingga .80 bagi dua keadaan pemarkahan. Manakala untuk reka bentuk kajian separa tersarang R: ($p \times T$) iaitu setiap pemeriksa akan menilai tugasan tertentu bagi pelajar tertentu, mencatatkan pekali G antara .38 hingga .52 (keadaan pemarkahan tunggal) dan antara .55 hingga .68 (dua keadaan pemarkahan) masing-masing. Dapatan kajian memaparkan pola yang ditunjukkan oleh bilangan keadaan pemarkahan yang berlainan pada asasnya adalah sama kecuali anggaran pekali G bagi keadaan pemarkahan kedua adalah lebih tinggi sedikit. Namun begitu, sekiranya perbezaan pekali G antara kedua-dua reka bentuk tersebut adalah sangat kecil, maka keadaan pemarkahan kedua adalah tidak berbaloi dijalankan berbanding dengan masa, kos dan usaha yang digunakan, misalnya bagi reka bentuk $p \times T \times R$ dan (R: p) $\times$ T.

Dapatan kajian menunjukkan pekali G bagi purata pemarkahan untuk setiap pelajar berdasarkan dua tugasan, sembilan pemeriksa dan dua keadaan pemarkahan dalam reka bentuk asal iaitu $p \times T \times R$ adalah .75. Menurut para penyelidik, reka bentuk ini tidak sesuai digunakan apabila bilangan pelajar dan tugasan yang terlibat adalah sangat besar. Oleh itu, operasi kajian D yang lain telah dijalankan untuk meninjau kesan penggunaan reka bentuk yang berlainan agar memastikan prosedur pemarkahan yang lebih boleh dipercayai diwujudkan. Dapatan kajian menunjukkan sumber yang agak banyak boleh dijimatkan dengan hanya kehilangan minimum dari segi *generalizability* berdasarkan penggunaan reka bentuk separa tersarang terutamanya reka bentuk $p \times (R{:}T)$. Kajian D juga menunjukkan perubahan bilangan karangan yang dinilai untuk setiap pelajar mempunyai pengaruh yang lebih besar daripada perubahan bilangan pemeriksa terhadap kebolehpercayaan skor

pemarkahan. Selain itu, dapatan kajian juga menunjukkan pertambahan penggunaan bilangan tugasan dan pemeriksa mempunyai kesan yang lebih besar dalam meningkatkan *generalizability* tentang pemarkahan berbanding dengan pertambahan penggunaan bilangan keadaan pemarkahan. Dapatan kajian juga menunjukkan sekurang-kurangnya empat tugasan dan pemeriksa diperlukan untuk memperoleh pekali G pada tahap .80 bagi reka bentuk $p \times T \times R$ dan $(R{:}\ p) \times T$ manakala reka bentuk
$p \times (R{:}T)$ hanya memerlukan bilangan tugasan dan pemeriksa sebanyak tiga sahaja untuk mencapai tujuan tersebut. Untuk reka bentuk $R{:}\ (p \times T)$, pekali G masih berada di bawah .70 walaupun empat tugasan dan pemeriksa digunakan.

Dalam satu kajian lain pula, Swartz et al. (1999) telah menjalankan dua kajian untuk meninjau anggaran kebolehpercayaan skor bagi pengukuran yang berlainan dalam penulisan karangan dengan menggunakan Teori G. Pengukuran yang berlainan tersebut termasuklah ujian pemiawaian, ujian buatan guru dan skema pemarkahan holistik dan analitik. Kajian-kajian tersebut juga cuba meneroka kemungkinan kesan terhadap anggaran kebolehpercayaan skor berhubung dengan penggunaan bilangan pemeriksa yang berlainan dan seterusnya menggunakan skor tersebut untuk membuat keputusan tertentu (keputusan relatif dan keputusan mutlak atau salah satu daripadanya). Reka bentuk kajian yang digunakan dalam kedua-dua kajian tersebut adalah reka bentuk tersilang satu faset, $p \times r$, yang melibatkan tiga sumber varian iaitu: (i) *persons* (p) atau pelajar yang menulis tugasan karangan; (ii) pemeriksa (r); dan (iii) interaksi pelajar × pemeriksa (pr).

Dalam kajian pertama Swartz et al. (1999), mereka menggunakan prosedur pemarkahan piawai iaitu subujian *Spontaneous Writing* mengenai *TOWL-2 Form B* (TOWL-2) untuk mengkaji kebergantungan skor karangan. Instrumen ini mengandungi lima subujian iaitu kematangan tema, kontektual kosa kata, kematangan sintaksis, kontektual ejaan dan kontektual gaya yang mempunyai nilai kebolehpercayaan antara pemeriksa yang tinggi iaitu tidak kurang daripada .95 kecuali subujian kematangan tema (.92) berdasarkan anggaran ujian klasik. Aspek-aspek subujian lain yang dikaji termasuklah *Spontaneous Writing Quotient* (SWQ) dan kebolehbacaan tulisan. SWQ merupakan gabungan skor berdasarkan skor daripada setiap subujian (tidak termasuk subujian kebolehbacaan tulisan). Manakala subujian kebolehbacaan tulisan adalah dipetik daripada TOWL asal memandangkan aspek penulisan ini dianggap penting oleh para pendidik.

Dalam kajian ini, sejumlah 251 orang pelajar sekolah menengah di kawasan pedalaman dan luar bandar di bahagian tengah Selatan Alabama yang berumur 11 hingga 15 tahun telah diminta menulis sebuah karangan berunsur naratif berdasarkan kehendak format *TOWL-2*. Mereka perlu mengarang satu cerita berpandukan rangsangan bergambar hitam putih yang mencerminkan satu ruang angkasa futuristik dalam masa 15 minit. Sebanyak 20 buah karangan naratif (lebih kurang 8%) telah dipilih secara rawak daripada kumpulan pelajar tersebut untuk digunakan dalam kajian ini. Empat orang pemeriksa yang terlibat semuanya mempunyai pengalaman mengajar dan pengalaman klinikal tentang TOWL-2. Setiap pemeriksa diberi tempoh 7 hingga 10 hari untuk menilai 20 buah karangan naratif tersebut setelah diberi program latihan pemarkahan yang cukup rapi.

Dapatan kajian G menunjukkan lebih tiga perempat daripada jumlah varian dalam skor bagi subujian TOWL-2 adalah disebabkan oleh perbezaan dalam kalangan pelajar iaitu objek pengukuran kecuali subujian kematangan tema (66.9%). Keputusan juga menunjukkan bahawa sumbangan σ^2 (*r*) dan σ^2(*pr*) adalah agak besar dalam subujian kematangan tema (19.4% dan 13.7%) dan kebolehbacaan tulisan (11.8% dan 23.3%). Manakala dapatan kajian D menunjukkan kebergantungan skor subujian TOWL-2 yang berlainan dipengaruhi oleh bilangan pemeriksa dan jenis keputusan yang dibuat dan impak magnitudnya adalah berlainan berdasarkan karangan tunggal berunsur naratif. Misalnya, apabila bilangan pemeriksa bertambah daripada seorang kepada empat orang, nilai pekali G bagi keputusan relatif dalam subujian kematangan tema (.83 kepada .95), kontektual gaya (.81 kepada .94) dan kebolehbacaan tulisan (.74 kepada .92) bertambah dengan mendadak manakala pertambahan tiga aspek subujian yang lain iaitu kontektual kosa kata (.98 kepada .99), kematangan sintaksis (.94 kepada .98) dan kontektual ejaan (.98 kepada .99) adalah lebih kecil. Sementara itu, nilai pekali G berdasarkan gabungan skor dalam subujian SWQ agak stabil iaitu bertambah daripada .97 kepada .98. Pola perubahan indeks kebergantungan atau pekali *phi* bagi membuat keputusan mutlak adalah agak sama dengan pekali G kecuali nilai pekali *phi* secara relatif adalah lebih kecil iaitu dalam lingkungan .65 hingga .99.

Dapatan kajian D menunjukkan penggunaan pemeriksa dan karangan tunggal tidak dapat memenuhi kriteria kebolehpercayaan skor Nunnally (1967) iaitu .90 untuk semua aspek subujian penulisan berdasarkan keputusan relatif. Kriteria tersebut dapat dipenuhi sekiranya dua hingga empat orang pemeriksa digunakan. Dalam hubungan ini, subujian kematangan tema memerlukan dua orang pemeriksa

(.91) manakala subujian kontektual gaya dan kebolehbacaan tulisan masing-masing memerlukan tiga (.93) dan empat (.92) orang pemeriksa untuk mencapai tahap tersebut. Ini juga menandakan pencapaian pelajar dalam subujian penulisan adalah tidak sama rata. Untuk meningkatkan kebolehpercayaan skor, para penyelidik berpendapat bahawa cara yang lebih berkesan dan menjimatkan adalah menerusi peningkatan mutu latihan atau menambahkan jumlah masa latihan pemeriksa atau kedua-duanya sekali. Pertambahan bilangan pemeriksa dianggap kurang sesuai kerana cara ini melibatkan kos yang besar.

Kajian kedua yang dijalankan oleh Swartz et al. (1999) adalah untuk menganggar impak tentang bilangan pemeriksa dan jenis keputusan yang dibuat terhadap pelajar berdasarkan kebolehpercayaan skor pemarkahan holistik dan analitik yang digunakan oleh guru dalam bilik darjah. Menurut mereka, prosedur pemarkahan berkenaan tidak mengikuti proses pemiawaian yang ketat. Namun begitu, prosedur pemarkahan tersebut banyak digunakan oleh pihak sekolah untuk membuat keputusan penting tentang kemahiran menulis pelajar dan kualiti pengajaran penulisan. Memandangkan hakikat tersebut, maka evidens untuk membuktikan skor daripada ujian berkenaan mempunyai takat kebolehpercayaan yang mencukupi adalah diperlukan.

Dalam kajian kedua ini, sampel karangan naratif 20 orang pelajar yang dipilih secara rawak daripada 42 orang pelajar sekolah rendah Gred 4 dan 5 di sekitar bandar bahagian tengah Utara Carolina telah dijadikan bahan kajian. Umur bagi 20 orang pelajar tersebut adalah antara 8.96 hingga 11.93 tahun (M=10.14, SD=.88). Pelajar tersebut diberi kebebasan untuk memilih dan menyiapkan salah sebuah cerita

daripada dua buah cerita dengan permulaan cerita diberikan. Seperti juga dalam kajian pertama, masa 15 minit diberikan untuk menyiapkan karangan tersebut. Tiga orang pemeriksa berpengalaman dan berkelayakan tertentu telah dipilih untuk menilai karangan pelajar tersebut. Para pemeriksa telah dilatih menggunakan pemarkahan holistik yang berskala 1 hingga 6 berdasarkan panduan pemarkahan naratif yang digunakan oleh badan *National Assessment of Educational Progress* (NAEP). Selain itu, mereka juga dilatih menggunakan rubrik pemarkahan analitik yang digubal oleh para penyelidik sendiri untuk mentaksir karangan cerita. Panduan pemarkahan NAEP memerlukan pemeriksa menilai bilangan, kualiti dan kesepaduan elemen cerita dalam karangan naratif. Manakala skala pemarkahan analitik yang digunakan mengandungi lapan dimensi yang berlainan. Dua dimensi mengenai struktur kekompleksan ayat (mudah dan kompleks), tiga dimensi berlainan mengenai jenis ralat ayat (sintaksis, semantik dan morfologi) dan tiga dimensi berlainan mengenai jenis ralat ejaan (salah mengeja kata akar, infleksi dan umum). Pada penghujung latihan, setiap pemeriksa diberi 20 buah karangan pelajar yang dipilih rawak dan diminta menyiapkan kerja pemarkahan dalam tempoh dua minggu. Dapatan kajian G menunjukkan amaun varian yang besar bagi skor holistik berpunca daripada pelajar (.95 atau 82.6%). Begitu juga keadaan bagi skor analitik kecuali $\sigma^2(pr)$ untuk ralat semantik dan infleksi.

Dapatan kajian menunjukkan apabila pemeriksa tunggal digunakan, pekali G bagi skor holistik ialah .83 manakala pekali G bagi skor analitik pula adalah daripada .45 hingga .90. Hanya satu skor analitik (ayat kompleks) dapat mencapai kriteria .90 yang ditetapkan oleh Nunnally (1967) iaitu .90. Sementara itu, apabila tiga orang pemeriksa digunakan, pekali G bagi skor holistik ialah .94 manakala bagi skor

analitik pula adalah daripada .71 hingga .97. Empat daripada lapan skor analitik tersebut dapat menepati kriteria .90. Skor yang tidak mencapai kriteria tersebut ialah ralat sintaksis, ralat semantik, kata akar dan infleksi. Menurut Swartz et al. (1999), keadaan ini mungkin disebabkan oleh aspek pemarkahan tersebut adalah lebih subjektif dan memerlukan penilaian yang lebih teliti daripada pemeriksa. Anggaran kebolehpercayaan bagi keputusan mutlak dengan menggunakan satu hingga tiga orang pemeriksa adalah sama atau hampir sama dengan nilai pekali yang ditunjukkan dalam keputusan relatif. Ini adalah kerana kesan ralat bagi pemeriksa adalah sifar atau hampir sifar. Kajian ini menunjukkan bahawa latihan lanjutan bagi pemeriksa adalah penting kerana pemeriksa yang juga merupakan guru kelas selalunya melibatkan diri dalam mentaksir penulisan pelajar dan membuat keputusan pengajaran bagi kelasnya. Para penyelidik juga mencadangkan untuk mencapai piawaian .90, unsur subjektif yang terdapat dalam skala pemarkahan bagi pelbagai aspek penulisan perlu diminimumkan.

Keputusan daripada kedua-dua kajian menunjukkan sebahagian besar varian bagi skor penulisan berpunca daripada $\sigma^2(p)$ atau perbezaan di kalangan palajar. Pemeriksa berpengalaman dan terlatih tidak menyumbang amaun varian yang signifikan kepada skor penulisan. Namun, amaun varian yang agak besar dalam skor adalah disebabkan oleh interaksi pelajar × pemeriksa (*pr*). Seperti yang dijangkakan, nilai pekali G cenderung meningkat apabila bilangan pemeriksa bertambah. Namun begitu, penggunaan spesifik tentang skor penulisan adalah pertimbangan penting bagi kebolehpercayaan skor tersebut. Misalnya dalam kajian pertama, kebolehpercayaan skor bagi semua subujian TOWL-2 termasuk juga *Spontaneous Writing Quotient* yang digunakan untuk membuat keputusan relatif adalah

memuaskan tanpa mengira bilangan pemeriksa yang digunakan. Bagaimanapun, ketika membuat keputusan mutlak, kebolehpercayaan skor untuk beberapa subujian TOWL-2 mencatatkan nilai di bawah kriteria .90 walaupun empat orang pemeriksa digunakan untuk menilai karangan.

Keputusan untuk kajian kedua yang melibatkan ujian bukan piawai pula menunjukkan kebolehpercayaan skor kedua-dua pemarkahan holistik dan analitik meningkat dengan bertambahnya bilangan pemeriksa. Apabila seorang pemeriksa digunakan, 8 daripada 9 dimensi penulisan yang dinilai adalah di bawah kriteria .90 untuk keputusan relatif dan mutlak. Tetapi, apabila 3 orang pemeriksa digunakan, 5 daripada 9 dimensi adalah melebihi .90 untuk keputusan relatif dan mutlak. Hasil keputusan ini penting kerana ia memaparkan masalah yang timbul apabila skor pentaksiran karangan berasaskan bilik darjah digunakan untuk membuat keputusan pengajaran dan dasar yang penting bagi pelajar.

Sementara itu, Liu dan Zhang (1998) telah menjalankan satu kajian tentang penggunaan teori G dalam pentaksiran penulisan karangan Bahasa Cina. Tujuannya adalah untuk mengkaji kesan pemeriksa dan kesan bentuk karangan dalam pentaksiran karangan dan seterusnya membuat analisis dan perbandingan bagi pelbagai kesan ralat yang diperoleh dalam kajian ini. Menurut mereka, dapatan kajian-kajian lepas menunjukkan varian bagi pemarkahan karangan adalah besar. Mereka juga menyatakan kelebihan teori G dalam pengungkaian varian skor mengikut sumber-sumber ralat tertentu dengan menggunakan analisis ANOVA, dan seterusnya memudahkan pengawalan varian dalam pentaksiran karangan.

Dalam kajian ini, seramai 20 orang pelajar sekolah menengah atas daripada sebuah sekolah di Beijing, China dipilih sebagai subjek kajian. Bahasa ibunda mereka ialah Bahasa Cina. Untuk memastikan subjek mewakili pelbagai tahap kebolehan menulis, rekod markah karangan subjek dalam lingkungan antara 50 markah hingga 90 markah ke atas (markah penuh 100) telah dipilih (Liu, 2001). Subjek dikehendaki menulis tiga buah karangan iaitu karangan bentuk penghujahan, naratif serta gabungan bentuk penghujahan dan naratif dalam tempoh sebulan. Skema penskoran yang disediakan adalah berpandukan penskoran pentaksiran karangan *Gaokao* (peperiksaan masuk ke institusi pengajian tinggi Negara China) dan kajian yang berkaitan tentangnya. Sampel karangan dinilai secara analitik berdasarkan tiga domain utama: kandungan, bahasa dan struktur / organisasi. Setiap domain diberi markah secara berasingan. Kesemua 60 sampel karangan telah dinilai oleh enam pemeriksa. Tiga daripada pemeriksa memiliki pengalaman mengajar Bahasa Cina melebihi tiga tahun manakala pemeriksa yang lain baru sahaja tamat latihan praktikal pengajaran penulisan karangan Bahasa Cina. Setiap pemeriksa perlu memarkah semua karangan yang diagihkan dan mesti mengikut ketertiban yang sama semasa pemarkahan.

Liu dan Zhang (1998) menggunakan reka bentuk tersilang dua faset iaitu $p \times t \times r$ di mana p ialah objek pengukuran yang mewakili kebolehan menulis pelajar, t mewakili faset tajuk karangan dan r mewakili faset pemeriksa dalam kajian G. Faset tajuk karangan dan pemeriksa adalah rawak. Data kajian ini dianalisis dengan menggunakan perisian GENOVA. Keputusan untuk setiap komponen varian dan peratus varian dalam kajian G adalah seperti berikut: p = 9.68 (7%), r = 48.80 (37%), t = 2.43 (2%), pr = 13.82 (11%), pt = 6.84 (5%), rt = 1.97 (2%), dan prt,e = 47.73

(36%). Dapatan kajian ini menunjukkan komponen varian benar untuk pelajar $\sigma^2(p)$ secara relatif adalah sangat kecil. Komponen varian untuk tajuk karangan $\sigma^2(t)$ yang kecil menunjukkan bahawa bentuk karangan yang berlainan tidak membawa perbezaan yang besar dalam pentaksiran kebolehan menulis pelajar. Sementara itu, faset pemeriksa merupakan penyumbang varian ralat yang paling besar.

Dalam kajian D, para penyelidik meninjau perkaitan perubahan antara bilangan pemeriksa dan karangan dengan kepersisan pengukuran berdasarkan reka bentuk tersilang dua faset $p \times T \times R$ kesan rawak. Dapatan kajian menunjukkan pekali G yang diperoleh adalah rendah sahaja iaitu.12 bagi tugasan karangan dan pemeriksa tunggal dan .57 apabila tiga tugasan karangan dan enam pemeriksa digunakan. Selain itu, mereka juga meninjau penetapan faset pemeriksa berdasarkan reka bentuk yang sama. Penetapan faset pemeriksa memberi impak yang lebih besar lagi kepada pekali G. Ini dapat dilihat apabila $r = 1$ dan $t = 1$, pekali G berubah dari .124 ke .301 iaitu meningkat sebanyak 127%; dan apabila $r = 2$ dan $t = 3$, pekali G akan berubah dari .361 ke .619 iaitu meningkat sebanyak 72%. Dapatan kajian mereka juga menunjukkan bahawa pertambahan bilangan pemeriksa akan memberi impak yang lebih besar kepada perubahan pekali G daripada pertambahan bilangan karangan.

Untuk meninjau sejauh mana pengaruh ralat penskoran ke atas bentuk karangan yang berlainan, pengkaji mengandaikan bentuk karangan tertentu dalam reka bentuk tersilang satu faset $p \times R$ dengan p mewakili pelajar dan R mewakili pemeriksa. Oleh sebab terdapat tiga bentuk karangan, maka sebanyak tiga reka bentuk tersilang satu faset $p \times R$ digunakan dalam kajian ini. Hasil kajian

menunjukkan bahawa karangan bentuk hujahan mempunyai pekali G yang paling rendah (.32) berbanding dengan gabungan bentuk penghujahan dan naratif (.45) dan bentuk naratif (.52). Ini bererti ralat penskoran untuk karangan bentuk penghujahan adalah paling besar. Jika dilihat dari segi faset pemeriksa, varian pemeriksa $\sigma^2(R)$ bagi karangan bentuk penghujahan adalah paling besar (51%) jika dibandingkan dengan bentuk naratif (35%) dan gabungan bentuk penghujahan dan naratif (33%). Juga varian benar pelajar $\sigma^2(p)$ bagi karangan bentuk penghujahan adalah paling kecil (6%) antara tiga bentuk karangan yang dikajikan. Berdasarkan dapatan kajian ini, Liu dan Zhang (1998) menyimpulkan bahawa pentaksiran karangan dengan menggunakan bentuk penghujahan (pekali G = .32) tidak dapat mengukur tahap kebolehan menulis pelajar dengan tepat berbanding dengan karangan bentuk naratif (pekali G = .52).

Liu dan Zhang (1998) juga menjalankan kajian untuk melihat ralat penskoran bagi setiap aspek pemarkahan iaitu kandungan, bahasa dan struktur / organisasi dengan menggunakan reka bentuk tersilang dua faset $p \times t \times r$ kesan rawak untuk menganggar varian setiap aspek tersebut secara berasingan. Dari segi aspek penskoran karangan ditinjau dari segi objek pengukuran *p,* peratusan varian keseluruhan bagi aspek kandungan adalah 3.3%, manakala untuk aspek bahasa dan struktur / organisasi adalah 6.7% dan 11.1% masing-masing. Ini menunjukkan pengukuran kebolehan menulis berdasarkan aspek struktur / organisasi adalah paling tepat manakala pengukuran dengan aspek kandungan adalah kurang tepat sekali. Jika dilihat dari segi faset pemeriksa *r*, peratusan varian keseluruhan aspek kandungan, bahasa dan struktur / organisasi ialah 33.6%, 35.6% dan 18.3% masing-masing. Ini menjelaskan variabiliti skor bagi aspek kandungan dan bahasa jauh lebih besar

daripada aspek struktur / organisasi. Sekiranya ditinjau dari segi komponen varian reja *pri,e*, peratusan varian keseluruhan bagi aspek kandungan (50%) dan struktur / organisasi (52.7%) adalah tidak jauh berbeza, bagaimanapun aspek bahasa (31.6%) nyata adalah lebih kecil. Ini menerangkan bahawa ralat rawak bagi aspek bahasa adalah lebih kecil daripada aspek kandungan dan struktur / organisasi.

Setelah membuat analisis, Liu dan Zhang (1998) mendapati faset pemeriksa merupakan sumber ralat yang paling besar dalam pemarkahan karangan. Justeru itu, mereka berpendapat pasukan pemeriksa yang mantap dan mengelakkan pertukaran pemeriksa yang kerap adalah penting untuk mengawal sumber ralat dan secara langsung meningkatkan kepersisan pengukuran dalam pemarkahan karangan. Pengekalan pasukan pemeriksa yang sama dalam penilaian bermakna faset pemeriksa adalah ditetapkan dan ini dapat meningkatkan kebolehpercayaan pemarkahan. Selain daripada itu, pemeriksa perlu menjalani latihan rapi sebelum melaksanakan tugas pemarkahan agar mereka dapat menguasai skema pemarkahan dengan cekap. Memandangkan sumbangan ralat penskoran dalam karangan bentuk penghujahan adalah paling besar, para penyelidik mencadangkan satu kajian perlu dijalankan untuk menggubal skema pemarkahan yang piawai dalam menilai karangan bentuk tersebut di samping setiap karangan harus dinilai oleh dua atau lebih pemeriksa supaya ralat penskoran dikuranganken. Dapatan kajian juga menunjukkan pengaruh aspek kebolehan menulis iaitu kandungan, bahasa dan struktur / organisasi adalah berbeza terhadap ralat penskoran. Antara aspek pemarkahan tersebut, aspek kandungan adalah paling tidak tepat manakala aspek struktur / organisasi adalah paling tepat. Ini bererti penggunaan aspek kandungan dalam menentukan mutu sesebuah karangan boleh dipertikaikan. Oleh itu, para

penyelidik juga mencadangkan agar nisbah wajaran bagi aspek kandungan dikurangkan dengan sewajarnya dan pada masa yang sama menambahkan nisbah wajaran bagi aspek struktur untuk penskoran karangan secara analitik demi meningkatkan kebolehpercayaan pengukuran karangan.

Dalam kajian ini, objek pengukuran hanya meliputi 7% daripada jumlah varian dan berdasarkan cerapan tunggal, pekali G yang diperoleh hanyalah .124. Manakala berdasarkan reka bentuk kajian asal ($\acute{n}_r = 6$, $\acute{n}_i = 3$), pekali G yang diperoleh ialah .57. Menurut Liu dan Zhang (1998), objek pengukuran masih kurang tepat. Namun begitu, ini mungkin disebabkan oleh kehomogenan sampel kajian kerana lebih homogen sesuatu sampel kajian lebih rendah pekali G yang diperoleh. Mereka menggunakan 20 orang pelajar sekolah menengah atas daripada sebuah sekolah di Beijing, China sebagai subjek kajian, ini menimbulkan pertanyaan tentang keperwakilan sampel iaitu sama ada sampel kajian tersebut dapat mewakili semesta cerapan kebolehan menulis pelajar atau tidak.

Mengenai kritikan ketidaksesuaian faset bentuk karangan dijadikan sebagai faset rawak dalam kajian G, Liu dan Zhang (2001) mendakwa bahawa tugasan karangan adalah dipilih secara rawak tetapi hanya mengambil kira tiga bentuk karangan tersebut. Tambahan pula, walaupun semesta faset bentuk karangan adalah terhad, masih boleh dilakukan pensampelan rawak. Sekiranya menetapkan faset bentuk karangan, maka ini akan bercanggah dengan amalan pentaksiran karangan yang sebenar iaitu tidak mengehadkan bentuk karangan yang ditulis dan diuji. Selain itu, penggunaan sesuatu bentuk karangan dalam kajian adalah bergantung kepada minat penyelidik dan tidak timbul persoalan sama ada sesuai atau tidak. Dari segi

kerangka kajian teori G, semesta cerapan bagi faset-faset pengukuran selalunya diandaikan sebagai rawak, tersilang dan daripada semesta yang mempunyai saiz sampel infiniti (Brennan, 1992). Namun, dalam kerja pentaksiran yang sebenar, keadaan ini adalah tidak mungkin dipenuhi. Andaian tersebut hanyalah bertujuan memudahkan kajian. Biasanya, apabila penyelidik merancang untuk membuat generalisasi keputusan kajian atau kajian D, faset pengukuran dalam semesta cerapan kajian G adalah diandaikan sebagai rawak.

Satu lagi kajian yang dijalankan oleh Lehmann (1990) menunjukkan *generalizability* tentang pengukuran pencapaian kemahiran menulis secara am merentas pelbagai tugasan yang digunakan. Beliau menggunakan data kajian *IEA International Study of Achievement in Written Composition* untuk komponen Jerman Barat yang melibatkan 1487 orang pelajar Gred 11 daripada 71 kelas dalam lapan jenis sistem sekolah yang berlainan di bandar *Hamburg*. Setiap pelajar dikehendaki menulis empat tugasan penulisan sama ada pendek atau panjang dengan memilih daripada sembilan rangsangan tugasan yang diberikan. Tugasan tersebut adalah tugasan naratif, penghujahan dan reflektif, surat nasihat dan analisis ulasan surat khabar serta *functional* di mana tugasan naratif dan *functional* diperuntukkan pilihan. Setiap tugasan karangan dinilai oleh dua orang pemeriksa terlatih dan bertauliah berdasarkan satu skala lima poin dengan kaedah *general merit*. Kaedah ini merupakan gabungan skor tanggapan keseluruhan dan analitik (kecuali mekanis dan kebolehbacaan tulisan). Pemeriksa adalah terdiri daripada guru yang merupakan penutur bahasa ibunda. Dalam kajian ini, tugasan karangan diagih-agihkan kepada lima orang pemeriksa dengan menggunakan reka bentuk bintang (*star design*) untuk

mendapatkan sepuluh kemungkinan kombinasi tentang pemeriksa agar kesan pemeriksa boleh dinilai dengan cara yang paling baik (Lehmann, 1993).

Lehmann melaporkan persetujuan jitu (exact *agreement*) dan persetujuan terdekat (*adjacent agreement*) merentas semua kombinasi pemeriksa adalah 73.2% dan 97.0% masing-masing. Manakala kebolehpercayaan antara pemeriksa ialah .885 dan kebolehpercayaan intra pemeriksa ialah .939 (daripada sejumlah 138 buah karangan yang dinilai sebanyak dua kali oleh semua pemeriksa *Hamburg*) bagi semua pemeriksa karangan. Dalam kajian ini, reka bentuk separa tersarang digunakan iaitu pemeriksa adalah tersarang dalam pelajar tetapi tersilang dengan jenis rangsangan penulisan yang boleh diwakili oleh $t \times (r{:}p)$.

Analisis Lehmann menunjukkan varian faset pelajar × tugasan (*pt*) adalah paling besar (57.2% daripada jumlah varian dalam skor) manakala varian pelajar (*p*) atau varian benar hanya meliputi 29.7% daripada jumlah varian. Beliau mendakwa amaun $\sigma^2(p)$ yang kecil berbanding dengan $\sigma^2(pt)$ akan membatasi usaha untuk mentafsir pencapaian penulisan di kalangan pelajar. Selain itu, beliau mengulas bahawa pemeriksa dapat menilai secara konsisten merentas karangan yang ditaksir walaupun rangsangan tugasan karangan adalah berlainan memandangkan faset tugasan (*t*) hanya meliputi 2.5%. Sementara itu, komponen varian $\sigma^2(r{:}p)$ iaitu gabungan kesan pemeriksa (*r*) dan interaksi pemeriksa × pelajar (*pr*) adalah kosong. Dalam kajian ini, model kesan rawak adalah lebih sesuai digunakan. Ini bermakna mana-mana empat tugasan tersebut adalah diandaikan berasal daripada semesta penulisan sekolah dan dinilai oleh mana-mana dua orang pemeriksa daripada para pengadil. Dapatan kajian menunjukkan berdasarkan dua orang pemeriksa dan empat

buah tugasan karangan, pekali G yang diperoleh ialah .646. Namun begitu, bagi pemeriksa dan tugasan karangan tunggal, pekali G hanyalah .30. Berdasarkan anggaran model tersebut, untuk mencapai pekali G yang memuaskan iaitu melebihi .85, maka sekurang-kurangnya 13 rangsangan tugasan adalah diperlukan. Sekiranya keempat-empat domain penulisan dianggap sebagai faset tetap, pekali G akan mencecah .957 berdasarkan dua orang pemeriksa dan empat buah tugasan karangan. Namun begitu, Lehmann mendakwa nilai pekali G tersebut telah diperbesarkan jika ditinjau dari segi kesahan dan andaian yang didasari dalam kajian tersebut.

Lane dan Sabers (1989) pula menggunakan teori G untuk mengkaji kebolehpercayaan skor 15 karangan bahasa Inggeris yang ditulis oleh pelajar merentas gred iaitu penutur asli berdasarkan sistem penskoran tertentu. Ujian karangan yang berdasarkan sistem penskoran tersebut mungkin membawa kepada pemantapan satu norma kebangsaan dalam mentaksir pencapaian penulisan pelajar di samping untuk mengesan pencapaian mutu karangan pelajar sama ada memerlukan bantuan pengajaran khas atau tidak. Mereka mendakwa instrumen pentaksiran penulisan tersebut adalah unik iaitu pemeriksa boleh menjalani latihan secara sendiri berdasarkan pakej arahan kendiri yang disertai dengan peraturan penskoran.

Dalam kajian ini, karangan untuk 15 orang pelajar daripada Gred 3 hingga Gred 8 dipilih secara rawak sebagai data kajian. Pelajar hanya diberi 10 minit untuk menghasilkan karangan tunggal iaitu menggambarkan perkara yang paling buruk dan paling baik tentang hari hujan. Seramai lapan orang pemeriksa yang mewakili berbagai-bagai bidang profesional dan tidak pernah melibatkan diri dalam kerja pemarkahan penulisan pelajar telah dipilih untuk melakukan kerja pemarkahan

penulisan. Mereka menjalani latihan pemarkahan terlebih dahulu sebelum memulakan kerja pemarkahan. Pemarkahan dijalankan dengan kaedah analitik. Markah diberikan berdasarkan empat kategori iaitu idea, perkembangan dan organisasi, struktur ayat dan mekanis atas satu skala tujuh mata bagi setiap kategori.

Pemeriksa perlu membiasakan diri dengan kriteria penskoran yang terdapat dalam empat kategori tersebut. Mereka membaca lima contoh karangan yang telah dimarkah berserta rasional pemberian markah berdasarkan kategori tersebut oleh pakar panel. Selepas itu, mereka menjalani latihan untuk memarkah lapan contoh karangan dan membandingkan markahnya dengan markah yang diberi oleh pakar. Sekiranya terdapat perbezaan lebih daripada satu poin, pemeriksa terpaksa merujuk balik kriteria penskoran. Arahan dalam pakej juga mencadangkan pemeriksa membuat keputusan awal tentang markah *anchor* dalam setiap kategori iaitu dua poin (rendah), empat poin (sederhana) atau enam poin (tinggi) yang mengandungi kriteria spesifik. Kemudian pemeriksa akan menentukan mutu karangan sama ada berada pada, di atas atau di bawah markah *anchor* untuk menjamin kejituan pemberian markah. Kemudian, meraka diminta menilai 15 sampel karangan pelajar secara bersendirian dengan jangkaan masa pemarkahan tiga minit untuk setiap karangan. Menurut Lane dan Sabers (1989), keadaan pemarkahan dan bilangan pemeriksa yang digunakan adalah bertujuan untuk membayangkan keadaan proses pentaksiran penulisan yang sebenar yang berlaku di sekolah.

Dalam kajian ini, data dianalisis dengan menggunakan reka bentuk tersilang dua faset kesan rawak iaitu pelajar (p) × pemeriksa (r) × kategori (k). Hasil dapatan dalam kajian G menunjukkan kesan pemeriksa adalah kecil (.102 atau 4.2% daripada

jumlah varian) berbanding dengan kesan pelajar, kesan pelajar × kategori dan kesan tiga hala yang berbaur dengan ralat rawak (31.6%, 24.0% dan 22.4% masing-masing). Dalam kajian D, reka bentuk gabungan (*mixed design*) dengan faset kategori pemarkahan ditetapkan manakala faset pemeriksa adalah rawak digunakan kerana penyelidik tidak berminat membuat genaralisasi melewati empat kategori tersebut. Kedua-dua pekali G dan indeks kebergantungan (pekali D) dilaporkan dalam kajian ini. Menurut para penyelidik, pekali G adalah penting bagi pemantapan interpretasi norma kebangsaan manakala pekali D adalah releven memandangkan skor esei diperlukan untuk mengenal pasti pelajar yang mungkin memerlukan bantuan pengajaran khas dari segi kemahiran menulis. Dapatan kajian tersebut menunjukkan berdasarkan satu hingga empat pemeriksa bergabung dengan karangan tunggal, pekali G yang didapati adalah antara .68 hingga .90 manakala pekali D adalah antara .62 hingga .87. Keputusan kajian D menunjukkan pekali yang tinggi akan diperoleh apabila tiga atau empat pemeriksa bergabung dengan karangan tunggal digunakan, hanya menerusi latihan pemarkahan secara kendiri yang mudah. Kajian tersebut juga menunjukkan pertambahan bilangan pemeriksa dari satu kepada dua telah memberi kesan dramatik dalam peningkatan kebergantungan skor (pekali G meningkat daripada .68 kepada .81; pekali D meningkat daripada .62 kepada .76), seperti dalam kajian lain (contohnya, Breland, Camp, Johns, Morris dan Rock, 1987; Bunch & Littlefair, 1988).

Dalam perkembangan yang lain, Bunch dan Littlefair (1988) telah menjalankan analisis *generalizability* terhadap dua bentuk karangan yang berbeza yang ditulis oleh 1000 orang pelajar Gred 9 dalam *Maryland Writing Test*. Mereka mengagihkan 24 orang pemeriksa terlatih untuk menilai sama ada 1000 buah

karangan naratif atau 1000 buah karangan pendedahan dengan menggunakan skala empat poin. Walaupun 12 orang pemeriksa diagihkan untuk menilai setiap satu rangsangan yang berlainan tetapi setiap karangan adalah dinilai sebanyak enam kali. Mereka melaporkan persetujuan jitu bagi pemeriksa merentas 15 kemungkinan kombinasi pemarkahan untuk karangan naratif dan pendedahan mencatatkan 78.8% dan 73.3% masing-masing manakala persetujuan terdekat bagi pemeriksa untuk kedua-dua bentuk karangan tersebut adalah 99.9%.

Dalam kajian ini, reka bentuk kajian dalam kajian G dan kajian D adalah sama iaitu pelajar tersilang dengan pemeriksa dan pemeriksa tersarang dalam rangsangan tugasan [$p \times (r{:}t)$ dan $p \times (R{:}T)$]. Dalam kajian G, penyelidik melaporkan komponen varian untuk faset pelajar × rangsangan karangan [$\sigma^2(pt)$] adalah paling besar iaitu 65.3% daripada jumlah varian manakala varian benar untuk pelajar [$\sigma^2(p)$] hanya meliputi 19.9% daripada jumlah varian. Justeru itu, penyelidik mencadangkan pemilihan rangsangan karangan harus dilakukan secara terperinci supaya pelajar tidak menerima skor yang berlainan berdasarkan rangsangan karangan yang berlainan. Dapatan kajian mereka menunjukkan apabila bilangan pemeriksa dan rangsangan karangan satu hingga empat digunakan dengan rangsangan karangan dianggap sebagai faset tetap, nilai pekali G dan pekali D yang diperoleh adalah sama iaitu daripada .91 hingga .99. Peningkatan pekali G yang paling besar akan berlaku apabila pemeriksa kedua ditambahkan tanpa mengira bilangan rangsangan karangan digunakan. Penyelidik juga menjalankan operasi kajian D yang melibatkan rangsangan karangan sebagai faset rawak. Kajian tersebut dijalankan untuk meninjau kebolehlaksanaan generalisasi dua rangsangan karangan yang berlainan tersebut untuk mewakili semesta ekadimensi yang lebih besar untuk rangsangan karangan.

Dapatan kajian menunjukkan pekali G untuk rangsangan karangan dan pemeriksa tunggal adalah .21 manakala pekali G untuk empat rangsangan karangan dan empat pemeriksa adalah .54. Memandangkan nilai pekali G yang rendah, penyelidik mendakwa bahawa usaha untuk mengeneralisasi keputusan kepada semua kemungkinan rangsangan karangan adalah tidak sesuai.

Manakala Breland, Camp, Johns, Morris dan Rock (1987) pula mengkaji pentaksiran kemahiran menulis bagi 267 pelajar menengah atas daripada enam buah institusi dengan menggunakan ujian aneka pilihan dan karangan. Dalam kajian tersebut, mereka meminta setiap pelajar menulis enam tajuk karangan yang berbeza bentuk atau mod iaitu bentuk naratif, pendedahan dan pemujukan dengan dua tajuk untuk setiap bentuk karangan. Karangan bentuk naratif dan pendedahan adalah dihasilkan dalam kelas. Pelajar diberi masa 45 minit untuk menyiapkan setiap tugasan tersebut. Bagaimanapun, tugasan pemujukan boleh didrafkan terlebih dahulu dan disiapkan luar daripada kelas. Untuk mengesan kemahiran menulis, mereka telah memadankan skor pemarkahan holistik berskala enam poin dan bilangan kesilapan dalam karangan yang berlainan bentuk dengan skor daripada enam jenis ujian aneka pilihan (contohnya, skor perbendaharaan kata dan bacaan daripada ujian verbal SAT), maklumat soal jawab, gred kursus dan pemarkahan kursus daripada pengajar.

Dalam kajian ini, reka bentuk kajian gabungan iaitu p x $(T : M)$ x $(R : M)$ telah digunakan di mana p merupakan pelajar yang menjawab keenam-enam tajuk karangan dalam tiga bentuk karangan yang berbeza, T dan M adalah tajuk dan mod karangan masing-masing. Dalam reka bentuk kajian ini, tajuk karangan dan pemeriksa adalah tersarang dalam bentuk karangan manakala pelajar adalah tersilang

dengan kedua-duanya dengan mod karangan adalah ditetapkan. Breland dan rakan-rakannya melaporkan anggaran korelasi bagi kebolehpercayaan pemeriksa untuk enam tajuk karangan seperti berikut: satu pemeriksa (.517 hungga .651), dua pemeriksa (.682 hungga .789) dan tiga pemeriksa (.762 hungga .848). Mereka juga melaporkan pekali G berdasarkan setiap tajuk dalam bentuk karangan tertentu (naratif, pendedahan dan pemujukan), dua tajuk karangan merentas bentuk karangan, dan bilangan satu hingga tiga pemeriksa untuk setiap tajuk karangan. Pekali G yang ditunjukkan adalah daripada .356 hingga .876 iaitu satu tajuk dalam satu bentuk karangan berdasarkan satu pemeriksa kepada enam tajuk merentas semua bentuk karangan berdasarkan tiga orang pemeriksa. Selain itu, mereka juga melaporkan pekali G untuk satu hingga tiga bentuk karangan, satu hingga tiga tajuk karangan berdasarkan setiap bentuk karangan, dan satu hingga empat pemeriksa. Seperti juga kajian-kajian lain, pertambahan pekali G paling besar berlaku dengan pertambahan pemeriksa yang kedua (misalnya, .42 kapada .53 dalam bentuk naratif) tanpa mengira bilangan bentuk karangan dan tajuk karangan yang digunakan.

Namun begitu, pertambahan pekali G yang lebih besar telah dicapai dalam kajian ini adalah dengan menambahkan tajuk karangan yang kedua sama ada dalam bentuk karangan yang sama (misalnya, .42 kepada .59 dalam bentuk naratif) atau merentas bentuk karangan (misalnya, .42 kepada .57 daripada bentuk naratif kepada pendedahan) tanpa mengira bilangan pemeriksa yang digunakan. Breland dan rakan-rakannya menyimpulkan bahawa untuk mencapai nilai tahap kebolehpercahayaan skor yang sama dengan ujian aneka pilihan yang piawai iaitu antara .85 hingga .95, maka sekurang-kurangnya empat buah karangan dan dua bentuk karangan yang berlainan diperlukan. Mereka juga menganggarkan bahawa untuk mencapai pekali G

pada tahap .80 dalam kajian ini, sekurang-kurangnya memerlukan enam buah karangan dan dua bentuk karangan masing-masing.

Secara ringkas, dapatan kajian-kajian lepas tentang penggunaan teori G dalam pentaksiran penulisan dapat dirumuskan seperti berikut:

Jadual 2.3

Rumusan Kajian-kajian Lepas Tentang Penggunaan Teori G Dalam Pentaksiran Penulisan

Kajian (Tahun)	Bil. Tugasan (Masa/Jenis)	Bil Pemeriksa (Semakan) utk satu tugasan	Kaedah Pemarkahan	Reka Bentuk Kajian	Kesan Model	Pekali
Chen, Niemi, Wang, Wang dan Mirocha (2007)	4 jawab 2 (3 karya sastera, 1 cerpen bukan fiksyen) (2 waktu kelas)	1– 4	Holistik (1– 4 poin)	2 faset tersilang p x T x R	Rawak	Pekali G: .37–.40
Brown (2007)	2 (135 minit / baca & tulis, naratif)	(1– 2)	Holistik (0 – 5 poin)	2 faset separa tersarang p x (R : T)	Rawak	Pekali G: .30 – .63
Schoonen (2005)	4 (1 pemujukan, 1 deskriptif & 2 pendedahan)	1– 5	Holistik (1– 5 poin)	2 faset separa tersarang p x (R : T)	Rawak	Pekali G: .32 – .40
			Analitik (2 aspek)			.21 – .30
Lee & Kantor (2005)	8 (Tugasan prototaip)	Fasa I (1– 2)	Holistik (1– 5 poin)	2 faset tersilang $p \times T \times R'$	Rawak	Pekali G: .44 – .88[a] .53 – .91[b]
				1 faset tersilang $p \times T$		.53 – .92
		Fasa II (1– 6)		2 faset separa tersarang R : (p x T)		.53 – .92[a] .59 – .94[b]
				(R : p) x T		.53 – .88[a] .62 – .92[b]
Sudweeks Reeve & Bradshaw (2004)	2 (2 keadaan pemarkahan)	1– 9	Holistik (9 poin)	2 faset tersilang $p \times T \times R$	Rawak	Pekali G: .34– .79[c] .39– .81[d]
				2 faset separa tersarang $p \times (R : T)$		.43– .83[c] .52– .86[d]
				(R : p) × T		.32– .78[c]

						.37– .80[d] .38– .52[c] .55– .68[d]
				R: ($p \times T$)		
Kajian (Tahun)	Bil. Tugasan (Masa/Jenis)	Bil Pemeriksa (Semakan) utk satu tugasan	Kaedah Pemarkahan	Reka Bentuk Kajian	Kesan Model	Pekali
Swartz et al. (1999)	1 -Ujian Piawai (TOWL-2) (15 minit/naratif)	1– 4	Holistik (gabungan skor bagi 5 subujian)	1 faset tersilang *p* x *R*	Rawak	Pekali G: .97 – .99 Pekali D: .94 – .98
			Analitik (6 dimensi)			Pekali G: .74 – .99 Pekali D: .65 – .99
	2 jawab 1 -Ujian Buatan Guru (15 minit / naratif)	1– 3	Holistik (1– 6 poin)	1 faset tersilang *p* x *R*		Pekali G: .83 – .94 Pekali D: .81 – .93
			Analitik (8 dimensi)			Pekali G: .45 – .97 Pekali D: .43 – .96
Liu & Zhang (1998)	3 (penghujahan, naratif & gabungan kedua-dua)	1– 6	Analitik (3 domain)	2 faset tersilang $p \times T \times R$	Rawak	Pekali G: .12 – .57
					Gabungan: Pemeriksa tetap & tugasan rawak	Pekali G: .30 – .62
				1 faset tersilang $p \times R$	Rawak	Pekali G: .32 – .52
Lehmann (1990)	9 jawab 4 (penghujahan & reflektif, surat nasihat, analisis ulasan surat khabar. Naratif & *functional* ada pilihan)	1– 2	*General Merit* (1– 5 poin)	2 faset separa tersarang *T* x *(R : p*)*	Rawak	Pekali G: .30 – .65
Lane & Sabers (1989)	1 (10 minit / rangsangan naratif)	1– 8	Analitik (4 kategori, setiap satu 7 poin)	2 faset tersilang *p* x *R* x *K**	Gabungan: Kategori tetap & pemeriksa rawak	Pekali G: .68 – .90 Pekali D: .62 – .87

Kajian (Tahun)	Bil. Tugasan (Masa/Jenis)	Bil Pemeriksa (Semakan) utk satu tugasan	Kaedah Pemarkahan	Reka Bentuk Kajian	Kesan Model	Pekali
Bunch & Littlefair (1988)	2 (naratif & pendedahan)	(1– 6)	Holistik (4 poin)	2 faset separa tersarang *p* x *(R : T*)*	Gabungan: Tugasan tetap & pemeriksa rawak	Pekali G: .91 – .99 Pekali D: .91 – .99
Breland, Camp, Johns, Morris & Rock (1987)	6 (45 minit / 2 naratif, 2 pendedahan, 2 pemujukan)	1– 3	Holistik (1– 6 poin)	2 faset separa tersarang *p* x *(T : M)* x *(R : M)*	Gabungan: bentuk karangan tetap, tugasan & pemeriksa rawak	Pekali G: .36 – .88

Nota. p = pelajar; R = pemeriksa; T = tugasan; T*= rangsangan tugasan ; M = mod atau bentuk karangan; [a]Berdasarkan satu pemarkahan per esei; [b]Berdasarkan dua pemarkahan per esei; [c]Berdasarkan keadaan pemarkahan tunggal; [d]Berdasarkan dua keadaan pemarkahan; K* = kategori; *General Merit* = gabungan skor tanggapan keseluruhan dan analitik (kecuali mekanis dan tulisan tangan); pekali G = pekali *generalizability*; pekali D = pekali *phi* = indeks kebergantungan.

Setelah menyorot kembali kajian-kajian lepas, beberapa kesimpulan boleh dibuat daripada kajian-kajian lepas tentang kebergantungan skor pentaksiran penulisan. Kebanyakan kajian ini meneroka pertambahan dalam *generalizability* skor berdasarkan bilangan pemeriksa (atau bilangan penilaian) dan bilangan tugasan karangan (misalnya, Breland, Camp Johns, Morris, & Rock, 1987; Brown, 2007; Bunch & Littlefair,1988; Chen, Niemi, Wang, Wang, & Mirocha, 2007; Lane & Sabers, 1989; Lee & Kantor, 2005). Pendekatan seperti ini adalah berkaitan dengan kerangka keberkesanan kos iaitu bilangan tugasan karangan dan pemeriksa (atau bilangan penilaian) yang optimum untuk mencapai tahap kebergantungan skor yang diingini atau menetapkan nilai pekali yang hendak dicapai secara *a priori*. Dari segi prosedur pemarkahan, kebanyakan kajian menggunakan pemarkahan holistik kecuali kajian Liu dan Zhang (1998), dan Lane dan Sabers (1989) yang menggunakan pemarkahan analitik. Manakala kajian Schoonen (2005) dan Swartz et al. (1999) pula

meninjau kebolehpercayaan skor daripada kedua-dua prosedur pemarkahan iaitu pemarkahan holistik dan analitik. Kajian Lehmann (1990) merupakan satu-satunya kajian yang menggunakan gabungan skor tanggapan keseluruhan dan analitik. Dari segi reka bentuk kajian, reka bentuk kajian yang mempunyai unsur tersarang adalah biasa digunakan, misalnya pemeriksa tersarang dalam bentuk karangan (Breland, Camp, Johns, Morris, & Rock, 1987; Bunch & Littlefair, 1988), tugasan karangan (Brown, 2007; Schoonen, 2005) dan penulis (Lee & Kantor, 2005; Sudweeks Reeve & Bradshaw, 2004). Reka bentuk kajian berunsur tersarang mempunyai kelebihan dari segi masa dan kos, dan sesuai digunakan dalam pemarkahan karangan bagi pentaksiran penulisan langsung. Semua kajian ini menganggap pemeriksa sebagai faset rawak kerana tujuannya adalah untuk menghasilkan skor yang boleh dipercayai menerusi mana-mana pemeriksa yang berkelayakan dan berpengalaman. Tiga kajian telah menetapkan tugasan sebagai faset tetap (Bunch & Littlefair, 1988; Lane & Sabers, 1989; Liu & Zhang, 1998) kerana ketiga-tiga kajian tersebut telah meninjau pentaksiran penulisan langsung dari perspektif pengendalian pengujian (*operational testing*). Manakala kajian Lane dan Sabers (1989) menggariskan kategori pemarkahan sebagai faset tetap kerana setiap kategori tersebut dianggap sebagai domain yang tersendiri.

Senario lain yang didapati daripada kajian-kajian tersebut boleh ditinjau dari segi keputusan dan kesimpulan mengenai *generalizability* skor hasil daripada pentaksiran penulisan. Seperti yang ditunjukkan, pertambahan kebolehpercayaan skor yang dramatik berlaku apabila bilangan pemeriksa atau penilaian kedua digunakan (misalnya, Brown, 2007; Lee & Kantor, 2005; Bunch & Littlefair, 1988). Selain itu, komponen varian yang paling besar biasanya melibatkan komponen-

komponen yang berkaitan dengan tugasan (Brown, 2007; Bunch & Littlefair, 1988; Chen, Niemi, Wang, Wang, Mirocha, 2007; Lee & Kantor, 2005; Schoonen, 2005; Sudweeks Reeve & Bradshaw, 2004). Justeru itu, pertambahan bilangan tugasan karangan biasanya mempunyai impak yang lebih besar terhadap kebolehpercayaan skor. Keputusan *generalizability* skor bagi pentaksiran penulisan juga biasanya sukar diramalkan dan ia bergantung kepada keadaan pentaksiran, pemarkahan pemeriksa dan kaedah pemarkahan.

2.3 Teori *Generalizability* (Teori G)

2.3.1 Kerangka Teori G

Tujuan asas ujian adalah untuk mendapatkan sampel kelakuan calon yang dicerap daripada sesuatu situasi ujian untuk digeneralisasikan kepada prestasi kelakuan calon yang ditunjukkan dalam situasi bukan ujian yang lain. Justeru itu, langkah-langkah perlu diambil untuk mengenal pasti sebarang pengaruh yang mungkin mengehad atau mengurangkan darjah kebolehan generalisasi sesuatu keputusan ujian. Dalam hubungan ini, cara menganggar secara kuantitatif kesan-kesan yang dihasilkan oleh sesuatu kondisi ujian tertentu menerusi keputusan ujian, dan setakat mana keputusan ujian boleh digeneralisasikan merupakan isu yang sentiasa diberi perhatian oleh pembina ujian dan pakar pengukuran.

Kemunculan teori G telah mencetuskan satu orientasi pemikiran baru untuk menyelesaikan masalah tersebut. Pada awal tahun 1960an, sekumpulan pakar pengukuran dari Amerika Syarikat yang dipelopori oleh L. J. Cronbach telah memperkenalkan secara rasmi teori pengukuran yang baru iaitu teori *Generalizability* (Cronbach, Gleser, Nanda, & Rajaratnam, 1972; Cronbach, Rajaratnam, & Gleser,

1963; Gleser, Cronbach, & Rajaratnam, 1965; Rajaratnam, Cronbach, & Gleser, 1963). Semenjak itu, teori tersebut semakin hari semakin mendapat perhatian. Perkembangan dan penggunaan teori tersebut juga bertambah luas dan mendalam berikutan kajian dan aplikasi ke atasnya oleh sarjana dari Amerika Syarikat dan beberapa buah negara di Eropah.

Pada asasnya, teori G dianggap sebagai lanjutan daripada teori ujian klasik (Nunnally & Bernstein, 1994; Feldt & Brennan, 1989). Teori G yang berasaskan teori ujian klasik telah membekalkan teori pengukuran satu kerangka konseptual yang luas serta satu set prosedur statistik yang kuat dan fleksibel untuk menangani isu-isu pengukuran secara sistematik (Brennan, 2001, p.2). Teori G bukan sahaja meminjam kaedah reka bentuk eksperimen daripada bidang psikologi tetapi juga menggunakan teknik statistik seperti analisis varian (ANOVA) dan analisis varian multivariate (MANOVA) sebagai asas bagi penganggaran varian dan menjalani kajian eksperimen. Namun begitu, menurut Brennan (2001), aspek yang paling penting dan ciri yang paling unik bagi teori G adalah berkaitan dengan kerangka konseptualnya. Konsep-konsep penting dan unik yang diutarakan, antara lain, termasuklah semesta cerapan yang boleh diterima dan kajian G serta semesta generalisasi dan kajian D.

Dari segi teori, sama ada teori ujian klasik atau teori G adalah berasal daripada teori persampelan rawak. Walaupun teori G membuka lembaran baru dari segi pengenalpastian kondisi pengukuran, pengungkaian varian ralat, andaian ujian setara secara rawak (*randomly parallel tests assumption*), reka bentuk pengukuran dan sebagainya, namun dilihat dari segi konsep asas dan kerangka teori pengukuran

dalam bidang psikologi, jelas menunjukkan bahawa kedua-dua teori berkenaan mempunyai hubungan yang rapat antara satu sama lain terutamanya dari aspek pewarisan dan kesinambungan perkembangan.

Dari segi kaedah penganggaran, teori G dalam kajian G pada asasnya menggunakan kaedah kesan rawak analisis varian dan analisis varian multivariate bagi mengungkaikan kesan varian dan kovarian. Berpandukan teknik statistik pinjaman tersebut, teori G barulah berupaya mengkuantitikan berbagai-bagai sumber varian ralat yang berpotensi mempengaruhi kebolehpercayaan skor dalam pengukuran, dan ini juga dapat memampasi kelemahan yang dihadapi oleh teori ujian klasik dalam pengungkaian sumber varian. Analisis varian, juga dikenali sebagai analisis variasi (Romanoski & Douglas, 2002), merupakan sejenis kaedah analisis variabel. Fungsi utamanya adalah untuk menganalisis amaun kandungan varian dalam sumber data eksperimen yang berlainan berbanding dengan varian keseluruhan, seterusnya menentukan takat pengaruh atau kesan variabel bersandar dan variabel tak bersandar dalam eksperimen. Kaedah ini biasanya digunakan untuk menangani variabel tak bersandar yang lebih daripada dua paras dalam kajian eksperimen.

Teori G pada dasarnya menggunakan min kuasa dua dan teknik pengungkaian analisis varian untuk mendapatkan anggaran komponen varian yang berkaitan dengan objek pengukuran, faset pengukuran yang berlainan dan interaksinya, dan interaksi objek pengukuran dengan faset pengukuran (Suen, 1990). Namun begitu, analisis varian dalam teori G tidak boleh disamakan dengan analisis varian dalam ujian statistik kerana teori G tidak perlu menjalankan ujian signifikan

bagi setiap kesan yang terlibat dan melaporkan nilai F manakala analisis varian perlu menjalankan ujian statistik (Brennan, 1983 & 2001). Kedua, teori G perlu mengenal pasti objek pengukuran dan faset pengukuran terlebih dahulu manakala analisis varian tidak diperuntukkan konsep tersebut untuk membezakan objek pengukuran dan faset pengukuran. Ketiga, dalam kajian psikologi secara umum, penyelidik biasanya mengharapkan kadar varian bagi kesan kaedah adalah lebih besar berbanding dengan varian keseluruhan, dan kadar varian bagi kesan calon (tahap calon yang berlainan menyebabkan keputusan yang berlainan) adalah kecil. Sementara dalam pendekatan teori G, penyelidik mengharapkan sumbangan kesan calon atau objek pengukuran merupakan sumber utama bagi varian, dan varian bagi kesan-kesan lain biarlah sekecil mungkin. Ini kerana varian yang merentasi objek pengukuran adalah varian benar (Suen, 1991, p. 46). Keempat, pada asasnya analisis varian adalah untuk meninjau sama ada terdapat perbezaan yang signifikan antara faktor eksperimen dengan keputusan eksperimen. Apabila memastikan bahawa kadar perbezaan bagi varian antara paras faktor adalah lebih besar berbanding dengan varian keseluruhan setelah menerusi ujian F, maka selesailah tugas bagi analisis varian. Manakala anggaran komponen varian yang dijalankan oleh teori G adalah bertujuan untuk kesediaan kajian keputusan (kajian D). Oleh itu selepas selesai penganggaran komponen varian (kajian G), maka seterusnya penyelidik akan menjalankan kajian bagi mendapatkan keputusan pengukuran yang dikehendaki iaitu kajian D. Kelima, pada lazimnya dalam kajian analisis varian atau kajian lain, kebanyakan kesan dalam reka bentuk tersarang tidak disebut sebagai kesan utama, misalnya kesan pemeriksa tersarang dalam tugasan ($r:t$) dalam reka bentuk $p \times (r:t)$. Tetapi teori G menyifatkannya sebagai kesan utama (Brennan, 2001, p.54). Malah, banyak reka bentuk yang biasanya ditemui dalam teori G sekiranya diberi

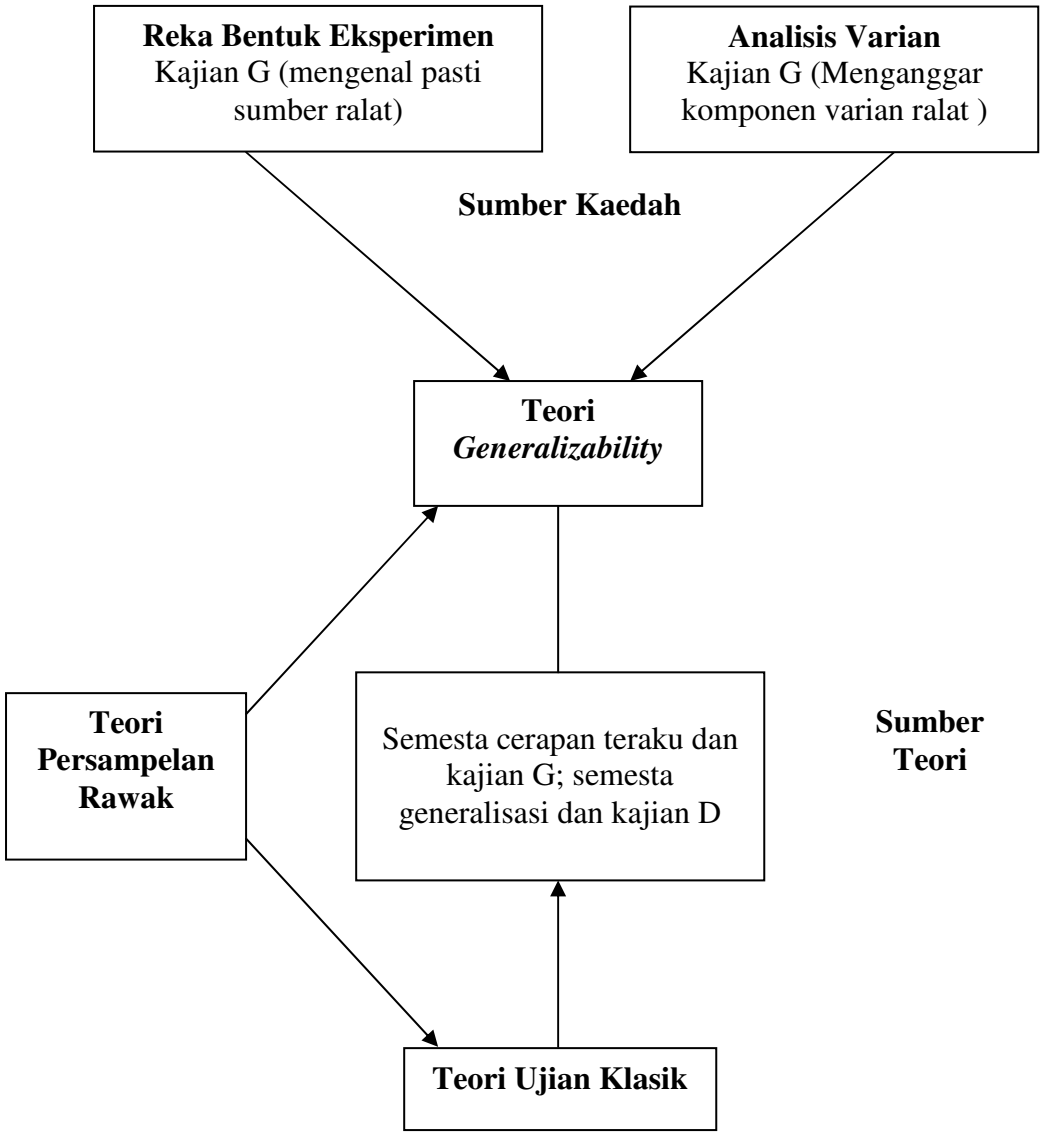
penerangan secara verbal untuk mencirikan reka bentuk tersebut mengikut konvensi ANOVA, maka sifat dan peranan faset dalam teori G akan menjadi keliru (Brennan, 2001, p.58).

Semasa menggunakan analisis varian, bilangan variabel atau bilangan paras variabel perlu diubahsuaikan menurut rancangan reka bentuk eksperimen yang berlainan. Begitu juga bagi keadaan teori G. Semasa mengungkaikan sumber varian yang pelbagai, teori G juga perlu mengikut struktur reka bentuk kajian yang ditetapkan. Justeru itu, reka bentuk eksperimen dalam bidang psikologi terutamanya pengertian reka bentuk eksperimen yang sempit, dijadikan salah satu panduan untuk pembentukan struktur reka bentuk teori G. Reka bentuk eksperimen dalam pengertian yang sempit pada asasnya melibatkan pelaksanaan rawatan eksperimen mengenai rancangan reka bentuk eksperimen dan analisis data statistik yang berkaitan. Antaranya, merancang reka bentuk eksperimen adalah bertujuan untuk mengawal ralat manakala menjalankan analisis statistik pula adalah untuk membuat kesimpulan-kesimpulan yang bermakna daripada data sampel. Kedua-dua cara tersebut mempunyai hubungan yang rapat kerana kaedah analisis statistik bergantung secara langsung pada reka bentuk yang digunakan. Teori G telah menerapkan pemikiran reka bentuk eksperimen tersebut ke dalam reka bentuk kajiannya dan menjadikannya sebagai panduan.

Namun begitu, ciri-ciri dalam reka bentuk eksperimen dan reka bentuk kajian teori G adalah berbeza. Pertama, langkah yang penting dalam reka bentuk eksperimen adalah mengenal pasti variabel tak bersandar, variabel bersandar dan variabel bebas (variabel yang perlu dikawal) dalam kajian (Mohd. Majid Konting,

1990, p.110). Berpandukan jalan pemikiran tersebut, teori G telah menunjukkan kemajuan tertentu. Teori G tidak mengikuti teori ujian klasik dengan menganggap ralat pengukuran sebagai ralat yang tidak dapat diasingkan (*undifferentiated error*) tanpa mempertimbangkan kondisi pengukuran yang konkrit. Sebaliknya teori G cuba mengesan semua faktor atau kondisi yang penting (mirip kepada variabel bebas dalam reka bentuk eksperimen) yang boleh mempengaruhi skor ujian. Dengan itu, keputusan pengukuran yang diperoleh boleh digeneralisasikan berdasarkan semesta kondisi yang konkrit. Kedua, reka bentuk eksperimen menggunakan tiga prinsip asas iaitu pereplikaan, perawakan dan pemblokan untuk mengawal variabel bebas yang memberi pengaruh kepada data eksperimen (Wu & Hamada, 2000). Perbezaan antara calon selalunya dianggap sebagai variabel bebas yakni penyelidik mengharapkan perbezaan calon adalah sekecil mungkin agar variabel eksperimen mampu memberi pengaruh maksimum terhadap data eksperimen. Sebaliknya, teori G biasanya mengharapkan kesan calon mampu menghasilkan pengaruh paling maksimum terhadap keputusan pengukuran manakala pengaruh kondisi-kondisi pengukuran iaitu faset-faset lain mempengaruhi skor ujian pada kadar yang paling minimum.

Untuk menyenangkan lagi pemahaman tentang seluruh kerangka teori G, sumber teori dan kaedah yang penting yang membawa kepada perkembangan teori G telah disimpulkan secara ringkas seperti dalam Rajah 2.2.



Rajah 2.2. Kerangka Teori G.

2.3.2 Perbandingan Teori G Dengan Teori Ujian Klasik

Teori G bukan sahaja berteraskan ciri-ciri istimewa teori ujian klasik, tetapi juga berjaya meminjam pemikiran mengenai kaedah reka bentuk eksperimen dalam bidang psikologi dan teknik analisis varian daripada bidang statistik. Dengan asas yang kukuh dan mantap, teori G berupaya merangkumi teori ujian klasik sebagai satu

kes khas dan melewati batas teori ujian klasik untuk menjelaskan keraguan konseptual di samping membekalkan satu instrumen statistik yang lebih kuat dan fleksibel.

Dalam teori ujian klasik, tiga komponen yang paling penting ialah andaian teori skor benar, kebolehpercayaan dan kesahan pengukuran. Sistem kaedahnya merangkumi teknik analisis item dan pemiawaian ujian. Manakala dalam teori G, objek pengukuran, faset pengukuran, semesta cerapan yang boleh diterima dan kajian G, semesta generalisasi dan kajian D, ralat relatif dan pekali G, ralat mutlak dan indeks kebergantungan adalah merupakan konsep penting dalam teori tersebut.

Walaupun teori G memperlihatkan ciri-ciri unik yang mampu mengambil alih kedudukan teori ujian klasik, namun begitu sehingga kini teori ujian klasik masih memainkan peranan penting dalam kerja pentaksiran secara praktikal. Ini kerana sistem teori dan kaedah ujian klasik bukan sahaja lebih mantap, malah premis bagi andaian model skor benarnya adalah lebih longgar dan mudah memenuhi syarat keadaan ujian yang sebenar terutamanya ujian skala besar. Model matematik, konsep parameter atau kaedah penganggaran dalam teori tersebut secara relatif adalah lebih mudah difahami dan digunakan. Walaupun begitu, teori klasik juga mempunyai kepincangan yang sukar diatasi. Dapatan pakar psikometrik yang menyemak semula teori klasik menunjukan bahawa andaian probabiliti yang mendasari pengukuran teori klasik sebenarnya melibatkan dua jenis ralat iaitu ralat rawak dan ralat sistematik (Thorkildsen, 2005, p. 68). Mereka juga membuat kenyataan bahawa semua pengukuran dalam sains sosial adalah mengandungi ralat dan ralat sistematik tetap merupakan komponen yang tidak dapat dikawal dalam sebarang skor.

Perbandingan teori G dengan teori ujian klasik adalah perlu dibuat memandangkan kedua-duanya mempunyai hubungan yang rapat dari segi kesinambungan teori di samping untuk memudahkan pemahaman teori G dengan lebih lanjut lagi.

2.3.3 Perbandingan Konsep Teori G Dengan Teori Ujian Klasik

Sama ada teori ujian klasik atau teori G mempunyai konsep mengenai objek pengukuran masing-masing yang agak berbeza. Dalam teori ujian klasik, objek pengukuran biasanya merujuk kepada sesuatu ciri terpendam (*latent trait*) seseorang yang secara relatif adalah lebih stabil dalam membuat inferens tentang suatu kebolehan sebenar seseorang. Sementara itu, teori G juga menganggap sesuatu ciri terpendam calon atau *persons* sebagai objek pengukuran, umpamanya kebolehan menulis calon. Namun, kadang-kala ciri-ciri tertentu bagi tugasan, pemeriksa, bentuk ujian, kumpulan kelas dan keadaan pemarkahan juga boleh dijadikan sebagai objek pengukuran berdasarkan minat penyelidik dan kepentingan kajian. Ini memandangkan teori G berkembang ekoran daripada latar belakang pemikiran yang berkaitan dengan pentaksiran prestasi dalam bidang pendidikan.

Menurut teori ujian klasik, skor cerapan calon boleh dibahagikan kepada dua bahagian iaitu skor benar yang merupakan sasaran proses pengukuran dan kombinasi ralat rawak yang tunggal (Bolus, Hinofotis, & Bailey, 1982; Brennan, 2001, p.2). Ia boleh digambarkan seperti berikut:

Skor Cerapan = Skor Benar + Skor Ralat

atau secara simbol,

$$X = T + E \qquad (1)$$

Skor benar boleh ditafsirkan sebagai min skor yang diperoleh calon dalam semua kondisi ujian yang mungkin (lihat Aiken & Groth-Marnat, 2006, p. 88; Lynch & McNamara, 1998). Min skor atau skor benar ini menyerupai konsep skor semesta dalam teori G. Bagaimanapun, teori G beranggapan bahawa ciri terpendam calon tidak boleh dinyatakan sebagai skor benar dengan kenyataan abstrak. Ia harus ditafsir dalam lingkungan kondisi tertentu berdasarkan keperluan reka bentuk dan keputusan pengukuran (Brennan, 1989 & 2001).

Dalam proses pengukuran, sekiranya faktor atau kondisi pengukuran yang berupaya mempengaruhi keputusan pengukuran adalah berlainan, maka takat inferens bagi keputusan juga akan berlainan. Jelaslah bahawa faktor atau kondisi pengukuran, yang dikenali sebagai faset, merupakan konsep unik bagi teori G. Misalnya penyelidik menjadikan kebolehan menulis calon sebagai objek pengukuran, dan semua faktor atau kondisi pengukuran yang boleh mempengaruhi skor ujian seperti tugasan karangan dan pemeriksa merupakan faset pengukuran. Setiap faset pula boleh dibahagikan kepada bilangan paras yang berlainan. Misalnya pengkaji menggunakan dua buah karangan yang berlainan bentuk untuk mengesan kebolehan menulis calon, maka faset tugasan karangan mengandungi dua paras. Demikian juga, jika tiga orang pemeriksa dipilih untuk membuat penilaian terhadap tugasan karangan, maka faset pemeriksa adalah tiga paras. Dengan itu, kesemua paras yang mungkin wujud bagi kedua-dua faset tersebut (bersaiz tak terhingga berdasarkan teori) merupakan semesta cerapan teraku (*universe of admissible observations*) masing-masing iaitu semesta cerapan bagi tugasan karangan dan semesta cerapan bagi pemeriksa. Setiap kali apabila sesuatu pengukuran dijalankan, ia merupakan satu sampel bagi semesta cerapan yang boleh diterima, dan setiap kali nilai cerapan yang didapati daripada pengukuran tersebut merupakan satu skor semesta. Ini juga

bermakna bagi setiap pengukuran yang dijalankan, nilai pekali kebolehpercayaannya adalah berbeza.

2.3.4 Perbandingan Konsep Kebolehpercayaan Teori G Dengan Teori Ujian Klasik

Dari perspektif penggunaan dan perkembangan ujian bahasa, konsep kebolehpercayaan sentiasa merupakan persoalan asas yang diberi perhatian dan pertimbangan. Dalam teori ujian klasik, terdapat tiga takrifan mengenai konsep kebolehpercayaan yang mempunyai nilai setara (Crocker & Algina, 1986; Dai, Zhang & Chen, 1999, p. 70). Takrifan pertama adalah seperti berikut:

$$r_{xx} = \sigma^2_t / \sigma^2_x \quad (2)$$

di mana r_{xx} mewakili kebolehpercayaan pengukuran, σ^2_t merujuk kepada varian skor benar dan σ^2_x menandakan jumlah varian atau varian skor cerapan. Secara konseptual, kebolehpercayaan adalah nisbah varian skor benar kepada varian skor cerapan bagi suatu kumpulan yang diuji (Bhasah Abu Bakar, 2003, p. 102; Thorkildsen, 2005, p. 69) seperti dalam persamaan (2).

Kedua, kebolehpercayaan adalah kuasa dua pekali korelasi bagi skor benar dan skor cerapan bagi suatu kumpulan yang diuji seperti dalam persamaan (3).

$$r_{xx} = \rho^2_{tx} \quad (3)$$

Takrifan ketiga iaitu kebolehpercayaan adalah pekali korelasi bagi satu ujian x (bentuk A) dengan mana-mana satu ujian setara klasik x' (bentuk B) seperti dalam persamaan (4).

$$r_{xx} = \rho_{xx'}$$

(4)

Takrifan pertama mengandaikan bahawa skor cerapan dan skor benar adalah dalam hubungan linear. Jumlah varian adalah hasil tambah antara varian skor benar dan varian ralat dan kedua-dua komponen tersebut adalah tidak berkaitan antara satu sama lain (Thorndike, 2005, p.113). Takrifan kedua adalah lanjutan daripada takrifan pertama. Kebolehpercayaan bagi takrifan pertama dan kedua adalah merujuk kepada data sekumpulan calon dan bukan menggunakan instrumen yang sama untuk melakukan ujian berulang terhadap calon yang sama seperti dalam takrifan ketiga. Dengan ini, tahap pengoperasian takrifan bagi takrifan pertama dan kedua adalah lebih tinggi. Namun begitu, kita tidak dapat mengetahui nilai skor benar bagi objek pengukuran berdasarkan kedua-dua takrifan tersebut. Oleh itu, mereka hanya mendukung maksud secara teoritikal. Sebaliknya, takrifan ketiga mempunyai maksud yang sebenar. Takrifan ketiga telah menerapkan konsep pemikiran ujian setara iaitu mengandaikan terdapat dua atau dua bahagian ujian yang mempunyai indeks kesukaran dan diskriminasi yang sama, dan menguji ciri terpendam yang sama. Anggaran korelasi antara kedua-dua ujian setara tersebut boleh menerangkan takat kebolehpercayaan sesuatu ujian.

Secara kesimpulan, penjelasan tentang teori skor benar dalam takrifan pertama dan kedua terhadap anggaran ralat pengukuran masih kabur. Ralat pengukuran hanya dirumuskan sebagai pencemaran terhadap kebolehpercayaan pengukuran sekali gus dan tidak dapat diungkaikan secara kuantiti. Oleh itu, jenis sumber ralat yang sebenar mempengaruhi skor benar tidak dapat dikesan. Pada hakikatnya, memang terdapat berbagai-bagai sumber ralat yang kompleks dalam

proses pengukuran sebenar. Maka adalah tidak munasabah bagi menggolongkan pelbagai sumber ralat kepada satu skor ralat yang tunggal. Untuk mengkaji persoalan kebolehpercayaan, penyelidik bukan sahaja perlu mengenal pasti dan menganggar sumber-sumber ralat yang boleh mempengaruhi keputusan pengukuran tetapi juga perlu mengambil ikhtiar untuk mengurangkan pengaruh variabel bebas dalam proses pengukuran (Bachman, 1990).

Teori ujian klasik menggunakan pelbagai kaedah untuk menentukan pekali kebolehpercayaan sesuatu ujian. Misalnya, kaedah ujian berulang (*Test-Retest Method*), kaedah bentuk selang seli (*Alternate Form Method*), kaedah belah dua (*Split Half Method*), pekali *Cronbach alpha*, KR-20 (*Kuder-Richardson* 20), KR-21 (*Kuder-Richardson* 21), pekali keserasian Kendall (W) dan sebagainya. Antaranya, pengiraan bagi tiga kaedah yang pertama adalah berasaskan pekali korelasi *Pearson Product Moment.* Pekali *Cronbach alpha*, KR-20 dan KR-21 digunakan terutamanya untuk menguji ketekalan dalaman ujian atau kehomogenan item ujian. Manakala pekali W biasanya digunakan untuk menganggar kebolehpercayaan antara pemeriksa yang lebih daripada dua orang pemeriksa.

Teori ujian klasik juga mengemukakan konsep kebolehpercayaan dalam pemeriksa, kebolehpercayaan antara pemeriksa dan kehomogenan antara pemeriksa. Namum begitu, teori ujian klasik hanya mampu mengesan satu variabel ralat setiap kali, ia tidak dapat menangani lebih daripada dua variabel secara serentak. Justeru itu, teori ujian klasik tidak dapat memberi maklumat yang lebih menyeluruh bagi skor yang diperoleh. Ini juga menyebabkan teori ujian klasik tidak dapat membuat inferens daripada skor yang diperoleh dalam sesuatu situasi kepada populasi yang

diingini. Keadaan ini berbeza bagi teori G. Teori G mampu menganggar ralat tertentu berdasarkan kondisi pengukuran yang berlainan. Ia bukan sahaja dapat mentafsir ralat dengan lebih spesifik malah mampu membuat inferens tentang variabel yang dikaji berdasarkan semesta generalisasi yang berlainan dan membuat generalisasi berdasarkan ruang lingkup yang sesuai. Teori G menganggap setiap faset pengukuran merupakan sumber ralat sistematik manakala sifat ketekalan objek pengukuran sendiri, dan semua kesan interaksi objek pengukuran dengan faset pengukuran digolongkan sebagai sumber ralat rawak.

Teori G mengelaskan ralat pengukuran kepada dua jenis iaitu ralat relatif yang biasanya diwakili oleh simbol δ dan ralat mutlak yang biasanya diwakili oleh simbol Δ. Ralat relatif dapat ditakrifkan sebagai perbezaan antara sisihan skor cerapan seseorang *person* (merujuk kepada calon) dengan sisihan skor semestanya (Brennan, 2001, p. 12). Ralat relatif merujuk kepada ralat pengukuran yang dihasilkan oleh semua ralat rawak dalam semesta generalisasi. Nilai ralat relatif boleh mempengaruhi kedudukan relatif seseorang calon dalam keseluruhan calon. Perbandingan tahap seseorang calon dengan keseluruhan calon semakin kurang tepat sekiranya nilai ralat relatif semakin besar. Manakala varian ralat relatif atau sisihan min kuasa dua [$\sigma^2(\delta)$] merupakan jumlah varian bagi semua kesan interaksi dengan objek pengukuran. Varian ralat relatif adalah bersamaan dengan varian ralat dalam teori klasik. Petunjuk ralat relatif dalam teori G digambarkan menerusi pekali G ($E\rho^2$). Pekali G adalah bersamaan dengan nisbah varian berkesan bagi objek pengukuran kepada hasil tambah varian berkesan bagi objek pengukuran dan varian ralat relatif. Ia adalah bersamaan dengan pekali kebolehpercayaan dalam teori ujian klasik.

Ralat mutlak pula ditakrifkan sebagai perbezaan antara skor cerapan seseorang calon dengan skor semestanya (Brennan, 2001, p. 11). Ralat mutlak meliputi ralat pengukuran yang disebabkan oleh semua kesan interaksi dan kesan utama faset dalam semesta generalisasi. Varian ralat mutlak [$\sigma^2(\Delta)$] merupakan jumlah varian bagi semua komponen varian kecuali komponen varian objek pengukuran. Petunjuk ralat mutlak dalam teori G digambarkan menerusi indeks kebergantungan (Φ). Indeks kebergantungan adalah bersamaan dengan nisbah varian berkesan bagi objek pengukuran kepada hasil tambah varian berkesan bagi objek pengukuran bercampur dengan varian ralat mutlak (Brennan, 2001, p.13).

Selain itu, dalam teori skor benar klasik, skor semesta calon didefinisikan sebagai purata bagi sebilangan besar pengukuran setara yang ketat. Varian skor benar merupakan varian bagi purata tersebut sementara konsep kebolehpercayaan ialah nisbah bagi varian skor cerapan kepada varian skor benar. Dalam Teori G, skor semesta calon didefinisikan sebagai purata bagi pengukuran dalam semesta generalisasi. Pengukuran tersebut tidak diandaikan sebagai pengukuran setara yang ketat (Crocker & Algina, 1986, p.159). Andaian ujian setara yang ketat dalam teori ujian klasik yang digambarkan dalam persamaan (4) adalah sukar dipenuhi dalam penggunaan sebenar kerana tidak mungkin bagi kita membina ujian yang betul-betul setara. Dalam hubungan ini, teori G menggunakan andaian yang secara relatif lebih longgar iaitu andaian ujian setara secara rawak yang lebih fleksibel dalam menangani persoalan pengukuran yang konkrit.

Teori G telah memaparkan perkembangan-perkembangan baru sama ada dari aspek kondisi pengukuran, pengungkaian varian ralat, andaian ujian setara secara rawak atau reka bentuk pengukuran dan aspek-aspek lain hasil daripada mewarisi asas teori ujian klasik. Namun begitu, pada hakikatnya semua aspek ini berlegar pada satu perkembangan penting dalam teras pemikiran pengukuran yakni reformasi dari segi konsep kebolehpercayaan. Crocker dan Algina (1986, p.158) semasa mengulas sumbangan penyelidik-penyelidik teori G pada peringkat awal berpendapat bahawa sumbangan mereka bukan setakat mengembangkan satu formula baru mengenai kebolehpercayaan tetapi telah mengembangkan satu orientasi pemikiran baru mengenai kebolehpercayaan, dan memilih kaedah pekali kebolehpercayaan serta varian ralat yang paling sesuai dengan kondisi pengukuran sebenar. Secara ringkasnya, perbandingan konsep asas dan konsep kebolehpercayaan antara teori ujian klasik dengan teori G telah ditunjukkan dalam Jadual 2.3.

Jadual 2.4

Perbandingan Konsep Asas Dan Konsep Kebolehpercayaan Teori Ujian Klasik Dan Teori G

		Teori Ujian Klasik	Teori G
1.	Objek Pengukuran	Ciri terpendam bagi calon.	Ciri terpendam bagi calon. Boleh juga terdiri daripada ciri tertentu bagi item, pemeriksa, bentuk ujian, kumpulan kelas dan aspek-aspek lain.
2.	Gambaran teori tentang objek pengukuran	Skor Benar Ia dinyatakan secara abstrak sebagai skor benar (kebolehan sebenar calon tanpa terdapat sebarang ralat pengukuran).	Skor Semesta Skor bagi tahap ciri terpendam calon di bawah semesta kondisi pengukuran yang konkrit.
3.	Sumber ralat pengukuran	Instrumen Pengukuran Biasanya merupakan satu set item ujian sebagai rangsangan untuk mencungkil respons.	Faset Pengukuran Setiap faset pengukuran merupakan sumber ralat sistematik.
		Ralat Pengukuran Ralat yang disebabkan oleh calon sendiri dan proses pentadbiran ujian iaitu persekitaran fizikal, kesan ketua pemeriksa dan gangguan di luar dugaan dan sebagainya.	Objek Pengukuran Dan Kesan Interaksi Faset Sifat ketekalan objek pengukuran sendiri dan kesan interaksi faset merupakan sumber ralat rawak.
4.	Jenis ralat pengukuran	Ralat Rawak Ralat yang sukar dikawal yang disebabkan oleh faktor kebarangkalian yang tidak berkaitan dengan tujuan pengukuran.	Ralat Relatif Kombinasi interaksi setiap faset dengan objek pengukuran dalam semesta generalisasi.
		Ralat Sistematik Ralat yang mempunyai kesan sistematik disebabkan oleh faktor yang tidak	Ralat Mutlak Semua kesan interaksi dan kesan utama faset kecuali objek pengukuran dalam semesta generalisasi

		berkaitan dengan tujuan pengukuran.	
5.	Andaian teori	Andaian ujian setara yang ketat	Andaian ujian setara secara rawak (*randomly parallel tests assumption*)
6.	Gambaran indeks bagi Ralat	<u>Kebolehpercayaan</u> Takat ketekalan bagi hasil pengukuran. Terdapat tiga keadaan: a) $r_{xx} = \sigma^2_t / \sigma^2_x$ r_{xx} mewakili kebolehpercayaan pengukuran, σ^2_t merujuk kepada varian skor benar dan σ^2_x menandakan jumlah varian atau varian skor cerapan. b) $r_{xx} = \rho^2_{tx}$ Kebolehpercayaan adalah kuasa dua pekali korelasi bagi skor benar dan skor cerapan suatu kumpulan yang diuji. c) $r_{xx} = \rho_{xx'}$ Kebolehpercayaan adalah pekali korelasi bagi satu ujian x dengan mana-mana ujian setara klasik x'.	<u>Ujian Rujukan Norma</u> [Pekali G ($E\rho^2$)] $E\rho^2 = \sigma^2(p) / [\sigma^2(p) + \sigma^2(\delta)]$ Sekiranya bilangan semesta generalisasi sekian banyak, bilangan $\sigma^2(\delta)$ dan $E\rho^2$ juga sekian banyak. Teori G boleh mempunyai pekali G yang berlainan berdasarkan takat generalisasi keputusan ujian yang berlainan. <u>Ujian Rujukan Kriteria</u> [Indeks kebergantungan (Φ)] $\Phi = \sigma^2(p) / [\sigma^2(p) + \sigma^2(\Delta)]$ Varian ralat mutlak [$\sigma^2(\Delta)$] meliputi semua varian bagi ralat rawak dan ralat sistematik. Ia ditaksirkan sebagai nisbah varian skor bagi objek pengukuran sendiri kepada jumlah varian skor.

Gambaran mengenai perbezaan pengertian teori ujian klasik dan teori G juga boleh dibuat menerusi perbandingan indeks utama antara kedua-dua teori. Untuk menjelaskan keadaan tersebut dengan lebih konkrit, komponen-komponen varian dalam reka bentuk kajian ini iaitu reka bentuk separa tersarang $p \times (r{:}t)$ dijadikan

sebagai contoh perbincangan. Reka bentuk ini mempunyai lima komponen varian iaitu $\sigma^2(p)$, $\sigma^2(t)$, $\sigma^2(r{:}t)$, $\sigma^2(pt)$ dan $\sigma^2(pr{:}t)$ (Jadual 2.4). Komponen-komponen varian tersebut menunjukkan keberubahan untuk semua sumber variasi berdasarkan cerapan tunggal pemeriksa dan tugasan daripada semesta cerapan pemeriksa dan tugasan. Simbol *p, r dan t* mewakili calon, pemeriksa dan tugasan.

Jadual 2.5

Komponen-Komponen Varian Dalam Reka Bentuk Separa Tersarang $p\times(r{:}t)$

Sumber-Sumber Variasi	Simbol Varian
Calon (p)	$\sigma^2(p)$
Tugasan (t)	$\sigma^2(t)$
Interaksi calon dan tugasan (pt)	$\sigma^2(pt)$
Pemeriksa tersarang dalam tugasan ($r{:}t$)	$\sigma^2(r{:}t)$
Reja ($pr{:}t$)	$\sigma^2(pr{:}t)$

Jadual 2.5 memaparkan perbandingan antara teori ujian klasik dengan teori G berdasarkan cara matematik. Perkara 1 menunjukkan varian skor semesta [$\sigma^2(p)$] dalam teori G adalah mirip kepada varian skor benar V(T) dalam teori ujian klasik. Sementara itu, varian ralat relatif [$\sigma^2(\delta)$] dalam teori G adalah bersamaan dengan anggaran varian ralat dalam teori klasik. Namun begitu, teori G mempunyai kelebihan dalam menganggar setiap sumber varian berbanding dengan varian keseluruhan, seperti ditunjukkan dalam Perkara 3. Ini bermakna teori G dapat menangani pelbagai sumber varian ralat pada masa serentak. Manakala teori klasik hanya menggambarkannya sebagai ralat tunggal yang tidak dapat diasingkan. Oleh

itu, anggaran ralat pengukuran bagi faset-faset tertentu dalam kajian G boleh dikawal dalam kajian D. Misalnya, amaun varian atau ralat pengukuran yang besar bagi faset tugasan dan pemeriksa dalam kajian G boleh dikurangkan dengan menambahkan bilangan tugasan dan pemeriksa dalam kajian D yang masing-masing diwakili oleh $ń_t$ dan $ń_r$. Selain itu, pembuat keputusan juga boleh menyusun strategi tertentu berdasarkan amaun varian yang didapati dalam kajian G. Misalnya, sekiranya nilai varian bagi interaksi calon × tugasan [$\sigma^2(pt)$] adalah hampir kepada kosong, maka untuk tujuan kepersisan pengukuran adalah tidak perlu untuk menambahkan bilangan tugasan dalam kajian D memandangkan prestasi calon adalah konsisten daripada satu tugasan kepada tugasan yang lain. Walaupun teori ujian klasik berupaya memberi anggaran mengenai pengurangan ralat pengukuran menerusi formula *Spearman-Brown Prophecy*, namun anggaran formula ini hanya terhad kepada satu faktor bagi setiap masa. Ia tidak dapat menganggar lebih daripada satu faktor secara serentak.

Perkara 4 mewakili varian skor cerapan dalam teori klasik. Teori G menafsirkannya sebagai jangkaan varian skor cerapan memandangkan ia akan digunakan untuk menganggar amaun varian dalam kajian D berdasarkan sebarang pemadanan pemeriksa dengan tugasan. Seperti juga teori klasik dalam perkara 5, jangkaan varian skor cerapan [$ES^2(p)$] dalam Teori G mengandungi varian skor semesta [$\sigma^2(p)$] dan komponen varian ralat yang lain. Seperti juga dalam perkara 3, ralat tersebut meliputi ralat rawak dan ralat sistematik. Pengenalpastian ralat sistematik seperti faset tugasan dan faset pemeriksa dalam kajian G akan membolehkan pengawalannya dalam kajian D.

Pekali G dalam perkara 6 diperoleh dengan membahagikan varian skor semesta dengan komponen-komponen yang membentuk jangkaan varian skor cerapan. Pekali G mewakili takat di mana skor cerapan boleh digunakan untuk menganggar skor semesta. Indeks ini adalah menyerupai pekali kebolehpercayaan dalam teori klasik. Bagaimanapun, teori klasik perlu menggunakan pelbagai jenis indeks kebolehpercayaan untuk menentukan kadar varian skor benar yang hadir dalam pengukuran. Manakala pengguna teori G hanya perlu menggunakan satu indeks pekali kebolehpercayaan iaitu pekali G (atau pekali phi) untuk tujuaan tersebut. Secara eksplisit, pengukuran tersebut boleh digeneralisasikan kepada semua pemeriksa dan tugasan. Pekali G juga boleh menggantikan pekali kebolehpercayaan seperti kebolehpercayaan dalam pemeriksa dan kebolehpercayaan antara pemeriksa. Juga, apabila semakin banyak faset ditambahkan dalam reka bentuk pengukuran teori G, anggaran pekali G menjadi semakin efisien.

Apabila tujuan pengukuran adalah untuk membuat keputusan relatif, perbezaan min antara pemeriksa atau tugasan tidak akan menjejaskan keputusan pengukuran. Dalam keadaan ini, anggaran ralat yang tepat ialah ralat relatif (lihat perkara 3). Bagaimanapun, sekiranya objektif pengukuran adalah untuk menentukan calon yang memenuhi tahap kriteria tertentu iaitu membuat keputusan mutlak, faktor ketegasan pemeriksa dan aras kesukaran tugasan perlu dipertimbangkan. Oleh itu, teori G membezakan ralat relatif dan ralat mutlak (lihat perkara 7). Manakala teori klasik tidak membuat pembahagian tersebut.

Komponen-komponen yang lain yang tidak melibatkan calon iaitu tugasan [$\sigma^2(t)$] dan pemeriksa tersarang dalam tugasan [$\sigma^2(r{:}t)$], mewakili perbezaan min

sebenar dalam pelbagi paras bagi setiap faset. Misalnya, komponen varian bagi faset tugasan mewakili perbezaan dalam tugasan yang diberikan. Jika aras kesukaran bagi tugasan adalah hampir sama, maka komponen varian bagi tugasan akan menjadi kosong atau hampir kosong. Penjelasannya adalah sama bagi keadaan faset pemeriksa tersarang dalam tugasan.

Jadual 2.6

Perbandingan Teori Generalizabiliti Dan Teori Ujian Klasik

	Teori G		Teori Ujian Klasik
1.	Varian skor semesta: $\sigma^2 (p)$	1.	Varian skor benar: V(T)
2.	Varian ralat relatif: $\sigma^2 (\delta)$	2.	Varian ralat: V(E)
3.	Komponen varian ralat relatif: $\sigma^2 (\delta) = \sigma^2(pt) / \acute{n}_t + \sigma^2(pr:t) / \acute{n}_r\acute{n}_t$	3.	Komponen varian ralat: V(E) =V(E)
4.	Jangkaan varian skor cerapan: $ES^2(p)$	4.	Varian skor cerapan: V(X)
5.	Komponen-komponen jangkaan varian skor cerapan: $ES^2(p) = \sigma^2(p) + \sigma^2(pt) / \acute{n}_t + \sigma^2(pr:t) / \acute{n}_r\acute{n}_t$	5.	Komponen-komponen skor cerapan: V(X) = V(T) + V(E)
6.	Pekali Generalizabiliti: $E\rho^2 = \sigma^2(p) / [\sigma^2(p) + \sigma^2(pt) / \acute{n}_t + \sigma^2(pr:t) / \acute{n}_r\acute{n}_t]$	6.	Pekali Kebolehpercayaan: r^2_{XT} = V(T) / V(T) + V(E)
7.	Komponen varian ralat mutlak: $\sigma^2(\Delta) = \sigma^2(t) / \acute{n}_t + \sigma^2(r:t) / \acute{n}_r \acute{n}_t + \sigma^2 (pt) / \acute{n}_t + \sigma^2(pr:t) / \acute{n}_r\acute{n}_t$	7.	Ralat Mutlak: Sama seperti V(E) (kadang-kala membuat anggaran yang rendah)

Sumber: Penyesuaian daripada Bolus, R. E., Hinofotis, F. B., & Bailey, K. M. (1982, p. 251). An introduction to Generalizability Theory in second language research. *Language Learning, 32(2),* 245-258.

2.4 Kerangka Konsep Kajian

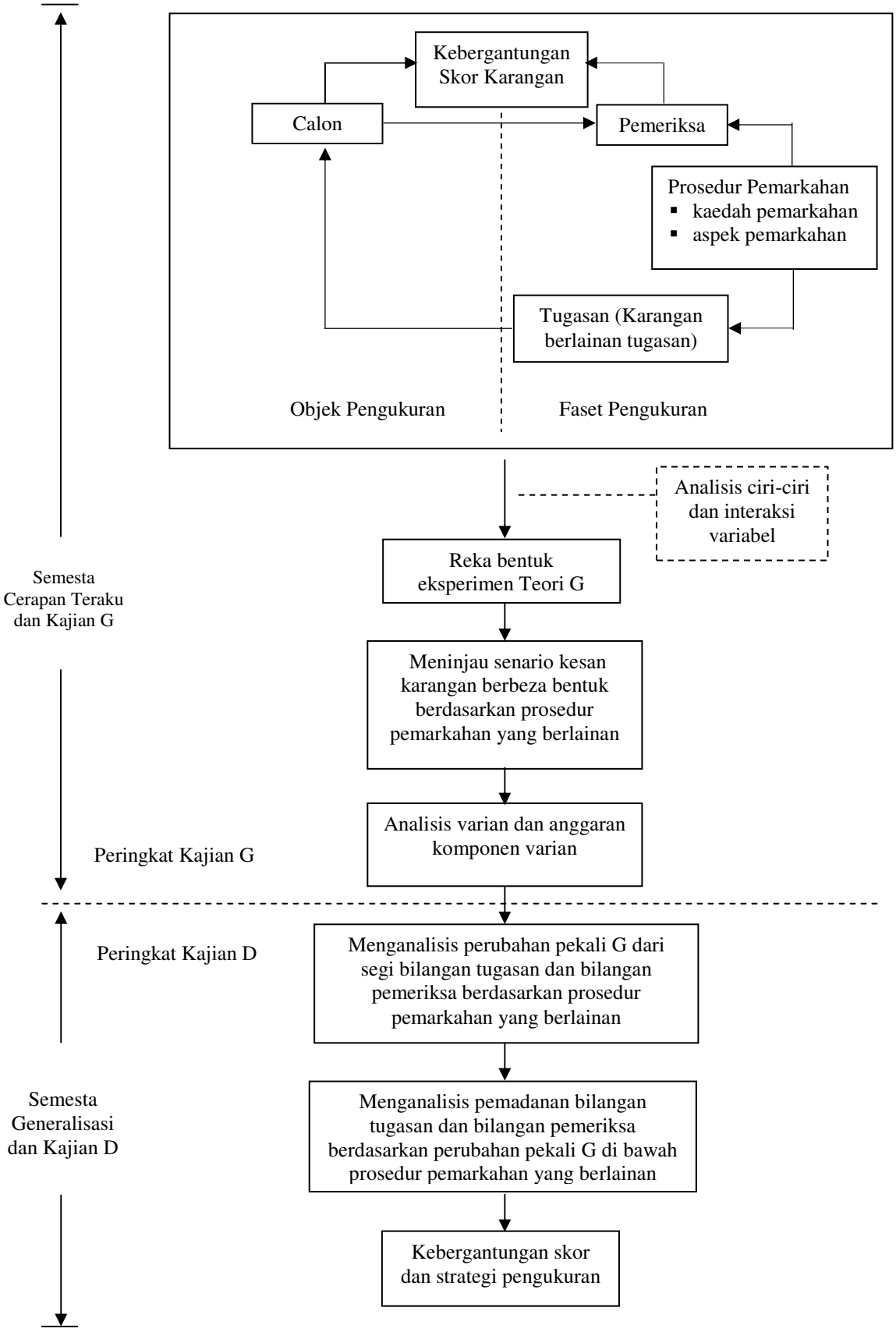
Kerangka konsep kajian menggambarkan penggunaan teori G untuk menganggar dan menganalisis pengaruh kesan pelbagai variabel dan juga interaksi

variabel-variabel tersebut terhadap skor karangan murid yang cuba diselidiki dalam kajian ini (Rajah 2.3). Kajian ini adalah untuk menyelidiki pengaruh variabel-variabel prosedur pemarkahan (kaedah pemarkahan dan aspek pemarkahan), karangan (berlainan tugasan) dan pemarkahan pemeriksa terhadap kebergantungan skor karangan murid.

Menurut istilah teori G, variabel calon merupakan objek pengukuran iaitu sumber varian benar. Manakala variabel-variabel prosedur pemarkahan, tugasan karangan dan pemeriksa dianggap sebagai faset pengukuran adalah sumber varian ralat. Secara terperinci, kajian ini bertujuan menyelidiki perkaitan antara kesan karangan berlainan tugasan dan kesan pemeriksa terhadap prosedur pemarkahan yang berlainan iaitu aspek pemarkahan dan kaedah pemarkahan berdasarkan kerangka kajian G, dan menganalisis impak gabungan bilangan tugasan dan bilangan pemeriksa bagi prosedur pemarkahan yang berlainan terhadap kebergantungan skor berdasarkan kerangka kajian D dari segi kebergantungan skor karangan berdasarkan teori G. Untuk menyelidiki perkaitan tersebut, pengkaji mula-mula menghuraikan ciri-ciri variabel tersebut menerusi keputusan min dan sisihan piawai berdasarkan statistik deskriptif. Manakala interaksi variabel-variabel tersebut iaitu interaksi karangan berlainan tugasan dengan prosedur pemarkahan yang berlainan diselidiki berdasarkan ujian perbezaan min dan kesamaan varian untuk mendapatkan gambaran awal tentang variabel-variabel tersebut.

Kajian ini menggunakan reka bentuk kajian teori G iaitu reka bentuk separa tersarang $p \times (r{:}t)$ untuk menganalisis kesan tugasan dan kesan pemeriksa

berdasarkan prosedur pemarkahan yang berlainan. Mengikut reka bentuk ini, pemeriksa adalah tersarang dalam karangan berlainan tugasan. Kajian ini meliputi



Rajah 2.3. Kerangka konsep kajian: Penggunaan teori G untuk menganalisis pengaruh kesan pelbagai variabel dan interaksi variabel-variabel tersebut terhadap skor karangan.

dua peringkat iaitu kajian G dan kajian D. Dalam kajian G, reka bentuk $p \times (r:t)$ digunakan untuk menganalisis kesan tugasan dan kesan pemeriksa berdasarkan terlibat. Kedua-dua kesan tersebut akan dianalisis berasaskan prosedur pemarkahan prosedur pemarkahan yang berlainan. Teknik analisis varian digunakan untuk mendapatkan anggaran komponen varian bagi kesan utama dan kesan interaksi yang yang berlainan. Seterusnya, kajian D dijalankan berdasarkan anggaran komponen varian dalam kajian G. Kajian D dijalankan untuk menganalisis perubahan pekali G berdasarkan bilangan paras tugasan karangan dan bilangan paras pemeriksa serta kombinasi optimum antara kedua-duanya berdasarkan prosedur pemarkahan yang berlainan. Selain itu, analisis perubahan pekali G berdasarkan prosedur pemarkahan yang berlainan juga dijalankan dengan memiawaikan faset tugasan.

2.5 Rumusan Bab

Dalam bab II ini telah dijelaskan pelbagai kajian berkaitan dengan faktor-faktor yang boleh mempengaruhi kesan tugasan karangan, pemeriksa, prosedur pemarkahan dan pengaplikasian teori G dalam pentaksiran karangan. Di samping itu, kerangka teori mengenai teori G dan kerangka konsep kajian yang digunakan juga dijelaskan.

BAB III

KAEDAH KAJIAN

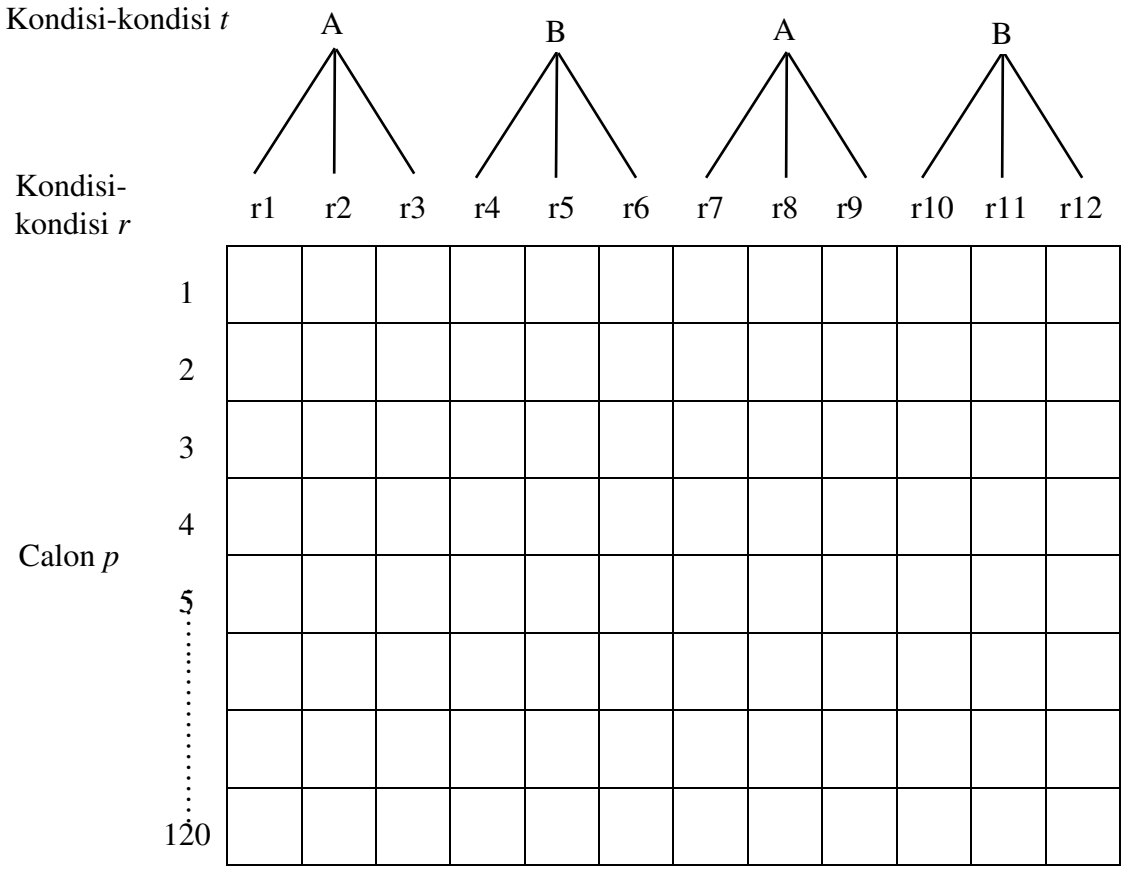
3.0 Pengenalan

Bab ini membincangkan kaedah yang digunakan dalam menjalankan kajian ini. Aspek-aspek penting yang disentuh termasuklah reka bentuk kajian, lokasi kajian, sampel kajian, instrumen kajian, tata cara kajian dan analisis data.

3.1 Reka Bentuk Kajian

Tujuan utama kajian ini adalah untuk mengkaji hubungan antara kesan tugasan karangan dan pemeriksa dengan prosedur pemarkahan yang berlainan ke atas kebergantungan skor berdasarkan penggunaan teknik statistik khas daripada teori G. Untuk menjalankan kajian ini, reka bentuk eksperimen dalam teori G iaitu reka bentuk dua faset separa tersarang $p \times (r : t)$ model rawak telah digunakan kerana kesesuaian dan kerelevanannya dengan tujuan kajian ini. Reka bentuk eksperimen ini memiliki kedua-dua kesan tersilang dan tersarang kerana ia merupakan gabungan daripada reka bentuk tersilang dan tersarang (Shavelson & Webb, 1991, p. 52). Dalam reka bentuk ini, objek pengukuran iaitu calon adalah diwakili oleh simbol *p* manakala simbol *t* dan *r* masing-masing melambangkan tugasan karangan dan pemeriksa, merupakan faset pengukuran dalam kajian ini di mana bilangan calon ialah 120 orang ($n_p = 120$) dan bilangan tugasan karangan ialah dua ($n_t = 2$) dan

setiap tugasan karangan akan dinilai oleh kumpulan pemeriksa (n_r = 3) yang berbeza (Rajah 3.1).



Rajah 3.1. Penggambaran reka bentuk dua faset separa tersarang $p \times (r{:}t)$ dalam bentuk skema.

Berdasarkan reka bentuk ini, pemeriksa tersarang dalam tugasan karangan yang berlainan mengikut kumpulan pemeriksa yang berbeza, dan kedua-dua faset pemeriksa dan tugasan adalah tersilang dengan calon iaitu setiap calon akan menjawab semua tugasan dan dinilai oleh setiap pemeriksa dalam kumpulan. Menerusi reka bentuk ini, kesan tugasan karangan dan kesan pemeriksa berdasarkan aspek dan kaedah pemarkahan yang berlainan dapat dianggar dengan teknik ANOVA. Kajian ini mengandaikan bahawa calon, tugasan karangan dan pemeriksa yang dipilih adalah rawak dan saiznya tak terhingga daripada semesta masing-

masing. Reka bentuk eksperimen ini telah digunakan dalam kerangka kajian G dan kajian D. Dalam kajian G, empat reka bentuk separa tersarang $p \times (r: t)$ model rawak berdasarkan empat prosedur pemarkahan yang berlainan telah dikendalikan. Anggaran komponen varian yang didapati daripada setiap reka bentuk eksperimen tersebut telah digunakan untuk membuat generalisasi bagi meninjau impak gabungan bilangan tugasan dan bilangan pemeriksa berdasarkan prosedur pemarkahan yang berlainan terhadap kebergantungan skor dalam kajian D. Reka bentuk seperti ini adalah sesuai digunakan untuk mengkaji kebergantungan skor karangan, misalnya dalam kajian *Explore Program* (ACT, 1994), analisis *Iowa Writing Assessment* (lihat Brennan, 1998, p. 323-325), *Manoa Writing Placement Examination* (MWPE) (Brown, 2007), kesan pemeriksa dan tugasan karangan terhadap *generalizalibity* skor penulisan (Schoonen, 2005), dan pentaksiran penulisan ESL baru (Lee & Kantor, 2005).

Selain itu, reka bentuk dua faset separa tersarang $p \times (r: t)$ model gabungan juga digunakan dalam kerangka kajian D. Dalam model gabungan ini, faset tugasan telah ditetapkan manakala faset pemeriksa kekal sebagai faset rawak. Dengan menetapkan faset tugasan, ini juga bermakna pengkaji tidak boleh menarik kesimpulan tentang apa yang akan berlaku jika tugasan yang berbeza digunakan. Dengan kata lain, semesta generalisasi untuk faset tugasan telah dihadkan. Secara kesimpulan, kepentingan reka bentuk eksperimen ini adalah digunakan untuk melihat impak gabungan bilangan tugasan dan pemeriksa berdasarkan prosedur pemarkahan yang berlainan terhadap kebergantungan skor dengan tidak membuat generalisasi ke atas tugasan karangan.

3.2 Lokasi Kajian

Kajian ini telah ditadbirkan di sebuah daerah di negeri Perak yang melibatkan sepuluh buah SJK(C) iaitu lima buah sekolah berada di kawasan bandar dan lima buah lagi di kawasan luar bandar. Sebanyak sepuluh buah kelas daripada sekolah-sekolah tersebut dipilih untuk menyertai kajian ini. Penggolongan sesebuah sekolah sebagai bandar atau luar bandar adalah ditetapkan oleh Jabatan Pelajaran Perak (JPP). Sekolah-sekolah yang berada di daerah tersebut merupakan sekolah pendidikan campuran dan hampir semua murid lelaki dan perempuan adalah terdiri daripada etnik Cina (lebih kurang 96%).

Daerah ini dipilih sebagai tempat kajian adalah berdasarkan keperwakilannya bagi daerah-daerah lain di negara ini yang mempunyai ciri populasi yang sama. Di samping itu, keadaan sedia ada bagi sekolah-sekolah di daerah tersebut mampu membekalkan data dan maklumat yang sesuai untuk tujuan kajian ini. Selain itu, sekolah-sekolah di daerah tersebut juga menunjukkan prestasi ujian penulisan yang agak konsisten untuk tempoh tiga tahun berturut-turut kebelakangan. Jadual 3.1 dengan jelas memaparkan pencapaian ujian penulisan Bahasa Cina dalam UPSR bagi murid Tahun Enam untuk sepuluh buah sekolah di daerah tersebut yang mempunyai bilangan murid sekurang-kurangnya lebih daripada 20 orang dalam Tahun 2005 hingga Tahun 2007. Pencapaian di atas tahap penguasaan minimum (peringkat A+B+C) ujian penulisan Bahasa Cina dalam UPSR bagi tiga tahun tersebut pada keseluruhannya adalah melebihi 80% kecuali SJK(C) L1 dan SJK(C) L3 dan purata keseluruhannya adalah lebih kurang 85%. Keputusan ini adalah hampir sama dengan

purata pencapaian kebangsaan (LPM, 2008). Secara analisis, peratus bilangan calon di luar bandar pada keseluruhan adalah kurang daripada 30% berbanding dengan calon kawasan bandar bagi Tahun 2005 hingga Tahun 2007 iaitu 28.2%, 26.9% dan 25.2% masing-masing.

Jadual 3.1

Ujian Pencapaian Sekolah Rendah (UPSR): Prestasi Ujian Penulisan Bahasa Cina Daripada Tahun 2005 Hingga Tahun 2007

Bil.	Nama Sekolah	Tahun	Bil. Calon	% Peringkat				
				A	B	C	A+B+C	D+E
1.	SJK(C) B1	2007	141	39.7	31.9	13.5	85.1	14.9
		2006	124	21.6	27.4	33.9	82.2	17.8
		2005	132	16.7	33.3	29.6	80.3	19.7
2.	SJK(C) B2	2007	244	27.5	47.1	18.4	93.0	7.0
		2006	254	37.2	41.5	14.2	92.9	7.1
		2005	242	31.4	33.5	19.8	84.7	15.3
3.	SJK(C) B3	2007	202	30.2	33.7	20.3	84.2	15.8
		2006	202	34.7	35.6	20.8	91.1	8.9
		2005	201	24.9	34.8	23.9	83.6	16.4
4.	SJK(C) B4	2007	240	50.8	34.2	12.1	97.1	2.9
		2006	302	41.7	42.3	11.6	95.7	4.3
		2005	291	50.2	38.1	7.9	96.2	3.8
5.	SJK(C) B5	2007	84	19.1	44.0	20.2	83.3	16.7
		2006	73	27.4	32.9	31.5	91.8	8.2
		2005	68	29.4	29.4	29.4	88.2	11.8
6.	SJK(C) L1	2007	55	10.9	40.0	20.0	70.9	29.1
		2006	73	19.2	16.4	31.5	67.1	32.9
		2005	73	24.7	30.1	13.7	68.5	31.5
7.	SJK(C) L2	2007	110	21.8	41.8	20.0	83.6	16.4
		2006	101	35.6	29.7	14.9	80.2	19.8
		2005	104	16.3	44.2	25.0	84.6	15.4
8.	SJK(C) L3	2007	92	20.7	27.2	23.9	71.8	28.2
		2006	123	23.6	24.4	22.8	69.7	30.3
		2005	126	27.0	37.3	15.1	79.4	20.6
9.	SJK(C) L4	2007	26	26.9	46.2	11.5	84.6	15.4
		2006	25	48.0	32.0	8.0	88.0	12.0
		2005	33	6.1	42.4	33.3	81.8	18.2
10.	SJK(C) L5	2007	24	16.7	58.3	16.7	91.7	8.3
		2006	30	40.0	40.0	10.0	90.0	10.0
		2005	30	50.0	46.7	0	96.7	3.3

Nota. Peringkat A, B dan C dianggap pencapaian di atas tahap penguasaan minimum manakala peringkat D dan E merupakan pencapaian di bawah tahap penguasaan minimum.
Simbol B = Bandar, L = Luar Bandar.

Sumber: Keputusan di atas didapati daripada pihak pentadbiran sekolah berkenaan masing-masing.

3.3 Sampel Kajian

Populasi kajian ini adalah terdiri daripada 1214 orang murid dalam 34 buah kelas Tahun Enam yang melibatkan10 buah SJK(C) di sebuah daerah di Perak. Sekolah-sekolah tersebut adalah di bawah penyeliaan dan pengawasan Kementerian Pelajaran Malaysia. Murid Tahun Enam dijadikan sebagai sasaran kajian kerana mereka dapat memberi satu gambaran keseluruhan tentang tahap pencapaian kemahiran menulis peringkat pendidikan sekolah rendah di SJK(C) sebelum melangkah masuk ke alam sekolah menengah. Manakala murid Tahun Enam di daerah tersebut dipilih sebagai sampel kajian adalah berdasarkan keperwakilannya dari segi pencapaian kemahiran menulis untuk murid Tahun Enam SJK(C) daripada daerah-daerah yang lain. Selain itu, mereka juga dapat memberikan maklumat yang lengkap dan diingini untuk kajian ini.

Dalam kajian ini, populasi murid dan bilangan kelas adalah berasaskan senarai yang didapati daripada pihak pentadbiran sekolah masing-masing di daerah tersebut pada awal tahun 2008 (lihat Jadual 3.2). Peratus populasi murid lelaki dan perempuan adalah agak sama iaitu masing-masing 50.4% dan 49.6%. Pada keseluruhannya, terdapat 21 buah SJK(C) di daerah tersebut. Namun begitu, pengkaji hanya melibatkan 10 buah SJK(C) dalam kajian ini kerana bilangan murid bagi sekolah-sekolah yang lain adalah terlalu kecil.

Dalam kajian ini, pensampelan rawak dua peringkat telah digunakan untuk memilih sampel kajian iaitu satu prosedur yang menggabungkan pensampelan rawak

kelompok dengan pensampelan rawak individu (Fraenkel & Wallen, 2007, p. 98). Reka bentuk pensampelan ini didapati lebih praktikal, memenuhi ciri keperwakilan

Jadual 3.2

Bilangan Subjek Kajian Mengikut Sekolah

Sekolah	Bil. Murid Tahun Enam		Bil. Kelas	Bil. Kelas Terpilih	Subjek Kajian
	Lelaki	Perempuan			
SJK(C) B1	73	72	4	1	12
SJK(C) B2	111	111	6	2	24
SJK(C) B3	115	77	5	1	12
SJK(C) B4	121	125	6	2	24
SJK(C) B5	48	43	3	1	12
SJK(C) L1	25	38	2	1	12
SJK(C) L2	53	56	3	1	12
SJK(C) L3	43	53	3	1	12
SJK(C) L4	13	11	1	-	-
SJK(C) L5	10	16	1	-	-
Jumlah	612	602	34	10	120
	1214				

Nota. Simbol B = Bandar; L = Luar Bandar.

subjek dan dapat membekalkan data yang diperlukan dalam kajian ini. Untuk menjalankan pensampelan tersebut, pengkaji menjadikan setiap kelas bagi Tahun Enam dan bukan setiap sekolah berkenaan sebagai unit kelompok (Wiersma & Jurs, 2005, p. 305-306). Dalam penyelidikan pendidikan, penggunaan kelas yang padu dan utuh sebagai kelompok merupakan satu amalan biasa (Ary, Jacob, Razavieh, & Sorensen, 2006). Pensampelan rawak kelompok adalah berguna apabila ahli populasi dapat dikumpulkan mengikut bentuk unit secara tabii dan boleh diaplikasikan dengan mudah sebagai kelompok (Wiersma & Jurs, 2005). Walaupun pensampelan ini akan melibatkan semua ahli kelompok dalam populasi yang dipilih sebagai sampel, namun penglibatan bilangan kelompok yang terhad mungkin mengakibatkan ralat

pensampelan (Mohd. Majid Konting, 1990) dan menjejaskan keperwakilan sampel kajian. Justeru itu, pensampelan rawak kelompok (dalam unit kelas) yang diiringi oleh pensampelan rawak individu seterusnya boleh mengatasi isu keperwakilan yang timbul dalam kajian ini.

Untuk keperluan tujuan kajian dan mengurangkan ralat pensampelan, pengkaji memilih secara rawak 10 buah kelas atau lebih kurang 30% daripada keseluruhan 34 buah kelas bagi Tahun Enam yang meliputi 10 buah sekolah di daerah tersebut. Kemudian, pengkaji melakukan lagi pemilihan secara rawak untuk mendapatkan lebih kurang 10% atau 120 orang murid sebagai sampel kajian. Ini dilakukan dengan memilih 12 orang murid daripada setiap kelas secara rawak (lihat Jadual 3.2). Pemilihan secara rawak ke atas murid-murid daripada 10 buah kelas tersebut adalah memastikan peluang yang sama bagi setiap murid memandangkan tahap kemahiran menulis mereka adalah pelbagai dalam setiap kelas. Namun begitu, semua murid daripada 10 buah kelas tersebut diminta untuk menjawab soalan yang disediakan. Kebolehlaksanaan prosedur ini didapati lebih tinggi berbanding dengan prosedur pensampelan rawak mudah, dan keperwakilannya adalah lebih memuaskan daripada pensampelan kelompok (Fraenkel & Wallen, 2007). Secara ringkas, pensampelan rawak dua peringkat dalam kajian ini melibatkan langkah-langkah seperti berukut:

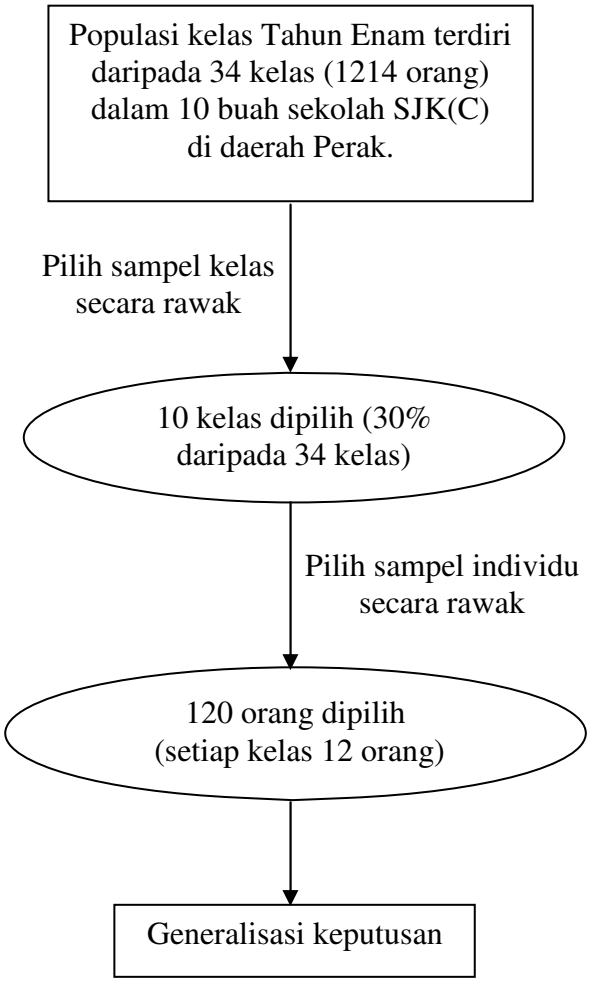
a) Populasi kajian yang dikenal pasti adalah 1214 orang murid Tahun Enam daripada 10 buah sekolah SJK(C) di daerah Perak.
b) Jumlah kelas yang terdapat dalam 10 buah sekolah tersebut ialah 34.

c) Saiz sampel kelas yang dipilih secara rawak ialah 10 kelas iaitu lebih kurang 30% daripada 34 kelas dengan menggunakan jadual angka rawak (Chua, 2006, p. 190-191).

d) Saiz sampel yang diingini dalam kajian ini ialah 120 orang murid Tahun Enam atau 10% daripada ahli populasi.

e) Pengkaji seterusnya memilih 12 orang murid daripada setiap kelas secara rawak berdasarkan 10 kelas berkenaan dengan menggunakan jadual angka rawak.

Carta aliran dalam Rajah 3.2 menunjukkan secara terperinci pemilihan sampel kajian dengan menggunakan kaedah pensampelan rawak dua peringkat.



Rajah 3.2. Pemilihan sampel dengan menggunakan pensampelan rawak dua peringkat.

3.4 Bahan Kajian

Sampel kajian diminta menulis dua buah karangan dalam Bahasa Cina sebagai bahan eksperimen untuk mentaksir kemahiran menulis mereka. Panjang karangan adalah tidak kurang daripada 120 patah perkataan. Karangan tersebut ditulis dalam jangka masa dua minggu. Pada minggu pertama, calon dikehendaki menulis sebuah karangan berunsur pendedahan dalam masa 30 minit. Selepas seminggu, calon menghasilkan lagi sebuah karangan berunsur naratif dengan peruntukan masa yang sama. Penulisan kedua-dua buah karangan tersebut adalah dalam keadaan terkawal iaitu menurut langkah-langkah ujian sebenar. Demi menjamin kebolehpercayaan dan kesahan skor tugasan karangan yang digunakan dalam kajian ini, prosedur-prosedur pengawalan mutu tugasan karangan telah dijalankan. Perkara ini akan dibincangkan dengan selanjutnya di bawah tajuk pemilihan tugasan karangan.

3.4.1 Takrifan kemahiran menulis

Menurut sukatan pelajaran KBSR Bahasa Cina, kemahiran menulis ditakrifkan sebagai kebolehan murid menulis karakter, kosa kata dan ayat serta menyampaikan idea menerusi pelbagai jenis tugasan penulisan yang berkaitan dengan ilmu pengetahuan dan pengalaman peribadi (KPM, 2003c). Penulisan karangan Bahasa Cina pada peringkat sekolah rendah adalah menekankan ayat yang gramatis, karakter dan tanda bacaan yang betul serta tulisan yang jelas dan kemas.

Di samping itu, murid juga digalakkan untuk menghasilkan karya penulisan berunsur ilmu dan imaginatif berdasarkan kreativiti mereka. Dari segi pengajaran kemahiran menulis pula, murid dihasratkan berkebolehan menghasilkan isi penulisan yang padat dan konkrit, tema jelas, penyampaian tersusun, ayat dan ungkapan yang lancar dan menarik serta penggunaan kosa kata dan karakter yang tepat (KPM, 2003b).

Sukatan pelajaran KBSR Bahasa Cina SJK(C) (KPM, 2003b, p. 15-16) telah menyenaraikan kandungan hasil pembelajaran untuk kemahiran menulis seperti berikut:

1. Menulis karangan berpandukan gambar dengan penyampaian yang teratur, menepati tema dan memerihalkannya secara konkrit.
2. Melengkapkan karangan dengan logik, kreatif dan teratur.
3. Menulis karangan bentuk naratif yang menarik, kreatif, jelas dan teratur.
4. Menulis karangan bentuk pendedahan secara objektif dengan ringkas, teratur dan jelas.
5. Menulis karangan bentuk pembahasan dengan memberi fakta, bukti dan membuat rumusan secara ringkas, tepat dan konkrit.
6. Menulis karangan berformat dengan betul, ringkas dan teratur.
7. Memindahkan maklumat daripada bentuk grafik kepada prosa.
8. Menulis puisi kanak-kanak dengan bahasa yang mudah dan kreatif.
9. Mengisi borang, menggubal kad dan borang berdasarkan keperluan keadaan dan aktiviti.

Jelaslah bahawa tugasan karangan yang harus dipelajari di peringkat sekolah rendah adalah berbagai-bagai. Dari segi pentaksiran karangan berasaskan sekolah, tugasan karangan yang lazim diuji adalah berbagai-bagai termasuklah karangan berdasarkan

gambar, melengkapkan karangan, karangan bentuk naratif, karangan bentuk pendedahan, karangan berformat, pemindahan maklumat dan karangan bentuk pembahasan. Manakala dalam pentaksiran karangan yang dikelolakan oleh peringkat daerah, negeri atau pusat pula, tugasan karangan berdasarkan gambar, karangan berformat, melengkapkan karangan, karangan pemindahan maklumat sama ada berunsur naratif atau pendedahan biasanya ditaksir sebagai mewakili kemahiran menulis murid. Biasanya tugasan karangan jenis pembahasan jarang sekali digunakan untuk mentaksir kemahiran menulis murid kerana tuntutan tugasan karangan tersebut kurang sesuai dengan perkembangan kognitif tahap murid Tahun Enam (Crowhurst, 1980).

3.4.2 Tugasan karangan

Tugasan boleh ditakrifkan sebagai suatu aktiviti seseorang itu melibatkan diri bagi tujuan untuk mencapai sesuatu objektif atau keputusan tertentu (Carroll, 1993). Untuk tujuan pengujian dan pentaksiran dalam bahasa, tugasan merupakan suatu aktiviti yang mempunyai konteks dan piawai, yang memerlukan pelajar menggunakan bahasa dengan menekankan makna dan berhubung dengan dunia sebenar, untuk mencapai sesuatu objektif, dan seterusnya mendapatkan data bagi tujuan pengukuran (Bygate, Skehan & Swain, 2001: 12). Secara spesifik, tugasan yang digunakan dalam pentakiran biasanya merujuk kepada item ujian yang dapat memperoleh prestasi yang kompleks dan kemahiran produktif secara langsung daripada calon (Davies, 2002). Dalam konteks pentaksiran karangan, tugasan merangkumi semua permintaan yang membolehkan calon untuk menyiapkan sebuah karangan pada akhirnya (Kroll, 1998).

Setiap tugasan karangan yang dihasilkan mempunyai tujuan tertentu iaitu sama ada untuk menceritakan atau melaporkan sesuatu, menggambarkan sesuatu, menerangkan atau memberi pendapat atau pandangan mengenai sesuatu, atau membincang dan membahas tentang sesuatu topik atau perkara. Sehubungan dengan ini, bentuk karangan mendukung tugas untuk menyampaikan isi atau mesej karangan tersebut. Dalam kajian ini, pengkaji menggunakan dua tugasan karangan iaitu karangan berunsur naratif dan pendedahan sebagai bahan kajian. Ini memandangkan hakikat bahawa karangan berunsur pendedahan dan naratif adalah asas kemahiran menulis dalam Bahasa Cina yang perlu dikuasai oleh murid-murid SJK(C) dan juga paling kerap dilatih dalam bilik darjah. Menurut Jacobs et al. (1981), bahan penilaian penulisan dikatakan mempunyai kesahan kandungan apabila jenis tugasan penulisan yang dinilai biasanya merupakan tugasan yang selalu dilatih dalam bilik darjah.

Karangan berunsur naratif merupakan jenis penulisan yang paling asas dan biasa didapati dalam bahan bacaan kanak-kanak. Karangan jenis ini biasanya mengkehendaki penulis menceritakan semula peristiwa-peristiwa yang telah berlaku berdasarkan urutan kronologi kejadian peristiwa. Untuk menarik minat pembaca dan kejelasan jalan cerita, penulis harus menampilkan masa, tempat, watak dan peristiwa dalam bahagian permulaan, perkembangan dan penutup karangan tersebut (Kamarudin Hj. Husin, 1993; KPM, 1997). Tugasan karangan ini merangkumi cerita, laporan, berita, rencana dan ulasan (Kamarudin Hj. Husin, 1988 &1993). Dalam peringkat sekolah rendah, guru bahasa biasanya melatih teknik penulisan naratif menerusi bentuk melengkapkan karangan. Dari segi penilaian juga, melengkapkan karangan merupakan bentuk yang lazim ditaksir. Dalam melengkapkan karangan,

jalan cerita memperlihatkan unsur logik, kreatif dan teratur (KPM, 2003b). Selain itu, bahasa penyampaian perlu jelas dan dapat menghidupkan suasana karangan.

Karangan berunsur pendedahan memerlukan penulis menerangkan sesuatu perkara dengan nyata agar pembaca boleh memahami dan mengenalinya dengan jelas. Oleh itu, penulis haruslah menulis isi yang tepat, fakta yang lengkap, penjelasan yang logik dan berkemampuan menyampaikannya dalam bahasa yang tepat, jelas dan tersusun. Selain itu, persembahan isi perlulah teratur dan menurut keutamaan serta mempunyai pemerengganan yang sesuai (KPM, 2003c). Karangan berunsur pendedahan juga adalah jenis karangan yang biasa untuk murid sekolah rendah di negara ini. Biasanya, murid berpeluang mempelajari teknik mengarang unsur penulisan ini menerusi mata pelajaran bahasa, mereka juga didedahkan dengan unsur penulisan ini melalui teks-teks pelajaran Sains, Sivik, Kajian Tempatan dan Pendidikan Moral. Ini jelas menunjukkan bahawa skop penggunaan karangan berunsur pendedahan adalah luas. Justeru itu, menguasai unsur penulisan ini adalah sangat penting terutamanya untuk murid sekolah rendah.

3.4.3 Pemilihan tugasan karangan

Tugasan karangan dalam kajian ini telah dipilih daripada soalan-soalan karangan Bahasa Cina tahun-tahun lepas yang terpakai dalam pentaksiran berasaskan daerah dan negeri di negeri Perak (Lampiran B). Tugasan karangan tersebut dibina oleh unit penggubalan soalan Bahasa Cina, Jabatan Pelajaran Perak (JPP) untuk tujuan pentaksiran pencapaian sekolah rendah. Prinsip pembinaan tugasan karangan adalah berteraskan kepada dua ciri utama iaitu kerelevanan dan keperwakilan. Dari aspek kerelevanan, tugasan karangan dibina berasaskan sukatan pelajaran dan

penekanan diberi kepada objektif sukatan pelajaran serta menepati kehendak jadual penentuan ujian. Soalan karangan yang dibina juga berkaitan dengan pengetahuan dan kemahiran yang pernah dipelajari oleh calon. Manakala dari aspek keperwakilan, unit tersebut juga memastikan bahawa tugasan karangan merupakan sampel secara rawak dari keseluruhan hasil pembelajaran yang dihasratkan yang perlu dikuasai oleh murid. Selain itu, karangan yang berlainan tugasan adalah dianggap sama dan diletakkan di bawah komponen kemahiran menulis dalam jadual penentuan ujian.

Untuk memastikan kesesuaian dan kualiti tugasan karangan dengan tujuan kajian, pengkaji telah mendapatkan bantuan empat orang ahli panel yang terdiri daripada guru-guru pakar Bahasa Cina dari SJK(C). Proses permohonan soalan karangan tahun-tahun lepas daripada JPN Perak dan proses mengundang panel penilaian soalan karangan telah ditunjukkan dalam Lampiran C manakala Lampiran D merupakan carta aliran yang berkaitan dengan prosedur kerja tersebut. Hampir semua tugasan karangan merupakan karangan berunsur pendedahan dan naratif. Ahli-ahli panel telah menilai kualiti karangan berunsur naratif dan pendedahan mengikut senarai semak yang disediakan (Lampiran E). Kriteria-kriteria penilaian dalam borang senarai semak mengandungi tiga aspek utama iaitu (a) kesejajaran dari segi kurikulum, (b) kejelasan dan ketepatan dari segi keseluruhan dan komponen tugasan karangan, dan (c) kesesuaian dari segi aras kesukaran dan keadilan. Kriteria-kriteria penilaian karangan digubal terutamanya berpandukan kriteria-kriteria penilaian kualiti item yang dikemukakan oleh badan peperiksaan dalam negara (LPM, 2003), Chatterji (2003, p. 219), Gronlund (2006, p. 124), Linn dan Gronlund (2000, p. 248) dan Mohamad Sahari Nordin (2002, p. 91). Selain itu, penggubalan kriteria untuk penilaian tajuk karangan juga mengambil kira skema analitik penilaian

tugasan karangan yang dikemukakan oleh Purves, Soter, Takala dan Vahapassi (1984). Kriteria-kriteria penilaian tentang aspek-aspek tugasan karangan yang disarankan merangkumi tuntutan kognitif, tujuan, peranan, sasaran pembaca, kandungan dan spesifikasi retorik. Skema berkenaan dibentuk untuk kajian penulisan karangan antarabangsa bagi kegunaan *International Association for the Evaluation of Educational Achievement* (IEA).

Sebelum memulakan kerja pemilihan tugasan karangan, pengkaji bertindak sebagai pengurus telah memberi taklimat ringkas mengenai objektif dan rasional kerja. Kemudian pengkaji mengelolakan latihan pemilihan tugasan karangan kepada empat ahli panel tersebut mengikut kriteria-kriteria yang disenaraikan berdasarkan beberapa buah tugasan karangan yang pilih secara khas. Setelah tamat latihan, setiap ahli panel diberi masa secukupnya untuk memilih lima tugasan karangan yang dianggap paling menepati kriteria-kriteria yang ditetapkan masing-masing daripada senarai 30 karangan berunsur naratif (N001-N030) dan 28 karangan berunsur pendedahan (P001-P028). Tugasan karangan yang diterima mestilah sekurang-kurangnya mendapat pilihan daripada tiga orang ahli panel. Sekiranya tidak dapat memenuhi syarat tersebut, pemilihan tersebut akan diulang semula sehingga mendapat bilangan karangan minimum iaitu tiga buah.

Dalam proses pemilihan karangan berunsur naratif, sebanyak empat tugasan karangan (diwakili oleh nombor siri) telah dipilih bersama oleh ahli-ahli panel iaitu N009, N012, N015 dan N028. Keputusan akhir pemilihan adalah seperti dalam Jadual 3.3. Proses yang sama dijalankan ke atas pemilihan karangan berunsur pendedahan. Dalam pemilihan karangan berunsur pendedahan, tiga tugasan karangan

telah dipilih bersama oleh ahli-ahli panel iaitu P013, P019 dan P021. Keputusan akhir adalah seperti dalam Jadual 3.4.

Jadual 3.3

Keputusan Pemilihan Karangan Berunsur Naratif

	Karangan berunsur naratif *	
Ahli Panel	Pilihan individu	Pilihan bersama
A	N003, N009, N012, N015, N022	N009, N012, N015 dan N028
B	N004, N009, N012, N015, N028	
C	N009, N011, N015, N026, N028	
D	N009, N012, N015, N022, N028	
Pemilihan secara rawak		N009

Nota. *Berdasarkan nomber siri.

Jadual 3.4

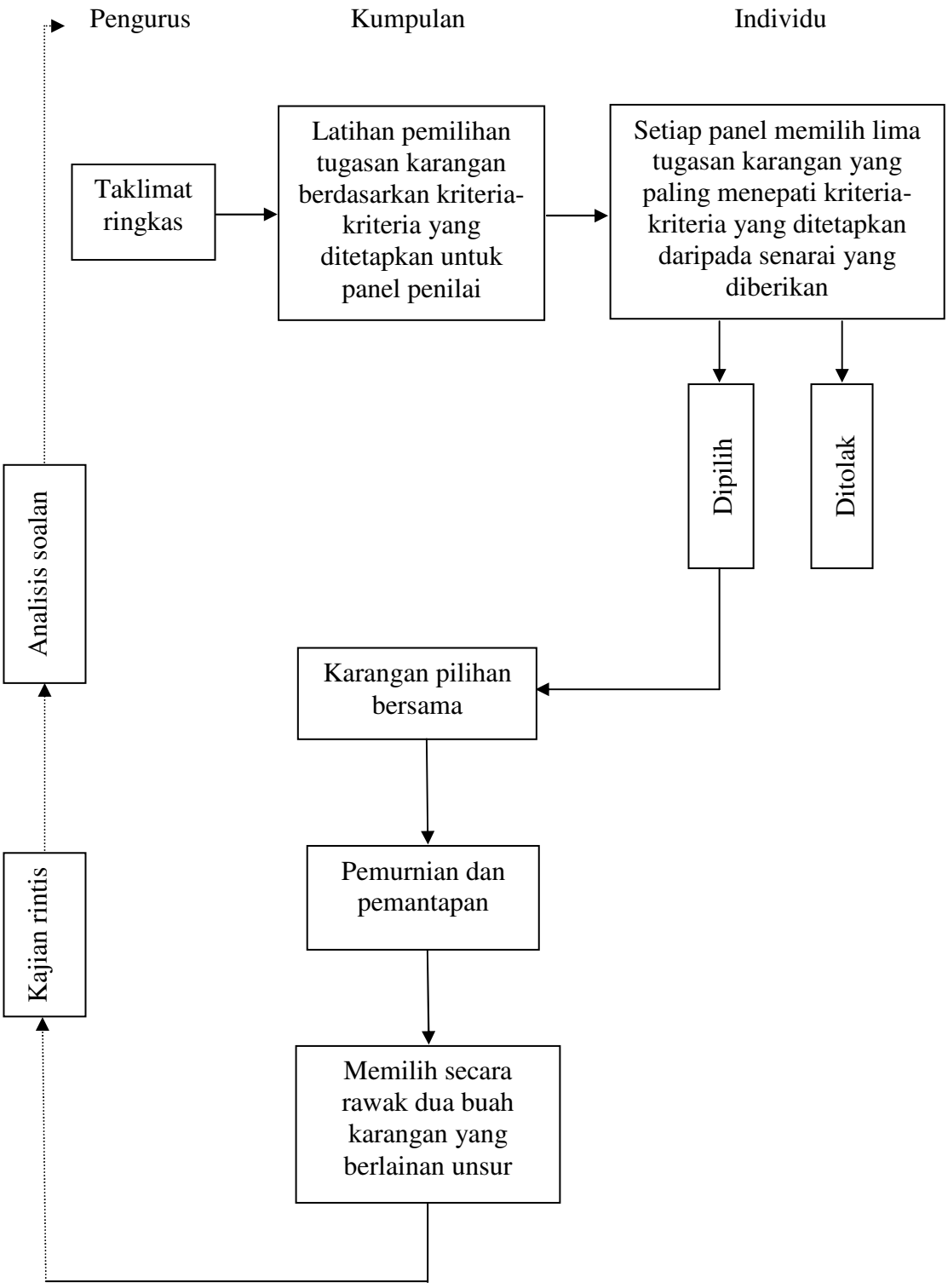
Keputusan Pemilihan Karangan Berunsur Pendedahan

	Karangan berunsur pendedahan*	
Ahli Panel	Pilihan individu	Pilihan bersama
A	P002, P012, P013, P019, P021	P013, P019 dan P021
B	P008, P013, P019, P021, P025	
C	P002, P011, P013, P019, P021	
D	P007, P013, P019, P021, P022	
Pemilihan secara rawak		P019

Nota. *Berdasarkan nomber siri.

Ahli-ahli panel seterusnya memurni dan memantapkan semua tugasan karangan yang telah dipilih agar mutu tugasan karangan terjamin sebelum dihantar untuk kajian rintis. Selepas pemurnian, pengkaji memilih secara rawak dua buah karangan untuk tujuan kajian rintis iaitu sebuah karangan berunsur naratif dan sebuah

lagi berunsur pendedahan. Tajuk karangan berunsur naratif iaitu dalam bentuk melengkapkan karangan memerlukan murid menceritakan hadiah yang diterima dengan menuntut sedikit daya imaginasi mereka manakala tajuk karangan berunsur pendedahan adalah berkaitan dengan penjagaan kebersihan sekolah. Kedua-dua tajuk karangan tersebut adalah mudah dan kurang membebankan tuntutan ingatan murid. Di samping itu, tajuk-tajuk tersebut membenarkan murid menumpukan perhatian tentang masalah dan perkara dalam dunia sebenar yang biasa dengan mereka menerusi pendedahan kepada aktiviti sekolah dan kehidupan harian. Tugasan karangan yang berlainan tersebut dianalisis selepas kajian rintis untuk mengenal pasti kesesuaian, kebolehlaksanaan dan kefahaman subjek terhadap tugasan karangan tersebut. Carta aliran dalam Rajah 3.3 menunjukkan proses pemilihan, pemurnian dan penganalisisan tugasan karangan.



Rajah 3.3. Carta penilaian tugasan karangan.

3.4.4 Pembentukan instrumen pemarkahan karangan

Dalam kajian ini, prosedur pemarkahan merujuk kepada aspek dan kaedah pemarkahan. Aspek pemarkahan merupakan hasil pembahagian kemahiran atau ciri tertentu tentang kemahiran menulis murid. Untuk tujuan kajian ini, aspek pemarkahan dibahagikan kepada dua komponen utama iaitu aspek kandungan dan organisasi serta aspek penggunaan bahasa dan mekanis. Kedua-dua aspek tersebut mempunyai ciri diskriminasi yang nyata dan juga berkait rapat antara satu sama lain (Schoonen, 2005). Linn dan Gronlund (2005, p. 240) juga beranggapan bahawa kemahiran menulis mengandungi dua kategori yang luas iaitu keberkesanan retorik dan konvensi atau kandungan dan mekanis walaupun dilihat dari perspektif skor analitik. Dalam kajian ini, setiap aspek pemarkahan akan dinilai dengan kaedah pemarkahan yang berlainan iaitu kaedah holistik dan analitik.

Untuk menilai karangan sampel kajian, instrumen pemarkahan yang mengandungi aspek dan kaedah pemarkahan yang berlainan telah dibina. Instrumen pemarkahan tersebut dibina berteraskan kepada beberapa sumber. Instrumen pemarkahan berdasarkan aspek pemarkahan yang berlainan dengan kaedah holistik dan analitik yang telah digunakan oleh LPM untuk mentaksir karangan Bahasa Cina peringkat UPSR (LPM, 2005) merupakan rujukan utama dalam pembinaan instrumen pemarkahan karangan untuk kajian ini. Di samping itu, ciri-ciri kemahiran menulis yang dinyatakan dalam kandungan sukatan pelajaran Bahasa Cina SJK(C) (KPM, 2003b) dan bahan pengajaran dan pembelajaran penulisan bahasa Cina KBSR

tahap II (KPM, 1997) daripada PPK juga merupakan sumber rujukan utama. Selain itu, aspek-aspek kemahiran menulis yang boleh dterapkan dalam pentaksiran karangan yang dikemukakan oleh Abdul Aziz Abdul Talib (1993) setelah meninjau pandangan pakar-pakar terkenal dalam bidang pengujian bahasa seperti Cooper (1977), Heaton (1979 & 1990), Oller (1979) dan Raimes (1983) juga dijadikan panduan untuk menggubal instrumen pemarkahan karangan untuk kajian ini. Mengikut Abdul Aziz Abdul Talib (1993, p. 162), lima aspek kemahiran menulis yang kerap digunakan dalam pentaksiran karangan adalah seperti berikut:

a. Kemahiran struktur dan gaya

   Kemahiran ini menekankan kebolehan menulis ayat yang betul, memanipulasi ayat dan menggunakan bahasa secara berkesan.

b. Kemahiran memilih isi

   Kemahiran ini menekankan kebolehan memilih isi karangan yang sesuai dengan tajuk.

c. Kemahiran perbendaharaan kata

   Kemahiran ini mencakupi kebolehan memilih, menggunakan perkataan yang beragam dan luas dengan tepat dan berkesan.

d. Kemahiran menyusun

   Kemahiran ini memerlukan kebolehan menyusun isi atau idea yang berkaitan secara teratur dalam turutan yang paling logik dan berkesan.

e. Kemahiran mekanis

   Kemahiran ini mencakupi kebolehan mengeja, menggunakan tanda bacaan, pemerengganan dan tulisan yang boleh dibaca.

Selain itu, pengkaji juga menyelidiki kajian yang dijalankan oleh pakar pentaksiran karangan Bahasa Cina dari negeri China iaitu Zhu (1990) mengenai aspek-aspek

pentaksiran karangan bagi murid sekolah rendah di negeri China. Aspek-aspek pentaksiran karangan yang disyorkan oleh Zhu iaitu aspek pemilihan isi (kandungan), organisasi dan bahasa (termasuk aspek retorik dan mekanis), adalah tidak jauh berbeza dengan senarai yang dikemukakan oleh Abdul Aziz Abdul Talib (1993).

Berpandukan sumber-sumber tersebut, pengkaji telah membentuk empat draf instrumen pemarkahan karangan. Keempat-empat draf instrumen pemarkahan yang dihasilkan adalah untuk memberi skor karangan berdasarkan prosedur pemarkahan seperti berikut: (a) kaedah holistik serta aspek kandungan dan organisasi, (b) kaedah analitik serta aspek kandungan dan organisasi, (c) kaedah holistik serta aspek penggunaan bahasa dan mekanis dan (d) kaedah analitik serta aspek penggunaan bahasa dan mekanis. Dalam kajian ini, aspek kandungan adalah merujuk kepada isi yang konkrit, padat dan sesuai dengan tajuk karangan manakala aspek organisasi merangkumi penyampaian yang jelas, tersusun, berkesan dan bertalian. Sementara itu aspek penggunaan bahasa meliputi kelancaran bahasa, struktur ayat yang lengkap, beragam dan tersusun serta kosa kata yang luas dan betul. Kriteria-kriteria yang ditekankan dalam aspek mekanis pula adalah berkaitan dengan penggunaan tanda bacaan dan karakter yang betul di samping kebolehbacaan tulisan.

Setiap instrumen pemarkahan diperuntukan 10 markah iaitu paling rendah satu markah dan paling tinggi sepuluh markah untuk menilai mutu karangan murid. Tidak ada wajaran yang lebih terhadap mana-mana satu aspek pemarkahan. Pembahagian markah yang sama rata untuk aspek-aspek kandungan, organisasi, penggunaan bahasa dan mekanis adalah berpandukan prinsip bahawa dalam

penghasilan karangan Bahasa Cina untuk peringkat sekolah rendah, aspek-aspek tersebut adalah tidak mudah diasingkan dan berkait rapat antara satu sama lain (KPM, 2003c; lihat Koh, 1981). Untuk pemarkahan kaedah holistik, kedua-dua aspek utama iaitu aspek kandungan dan organisasi serta aspek penggunaan bahasa dan mekanis dinilai secara keseluruhan. Manakala untuk pemarkahan kaedah analitik, aspek kandungan dan organisasi ditaksir secara berasingan, dan kemudian markah bagi kedua-dua bahagian tersebut dijumlahkan. Begitu juga cara pengiraan untuk aspek penggunaan bahasa dan mekanis dalam kaedah analitik.

Dalam kajian ini, kandungan keempat-empat draf instrumen pemarkahan dimurni dan dimantapkan oleh tiga orang guru pakar. Mereka merupakan guru cemerlang dan juga Jurulatih Utama PPK untuk mata pelajaran Bahasa Cina SJK(C). Pakar-pakar tersebut diberi taklimat terlebih dahulu sebelum memurni dan memantapkan draf-draf berkenaan berdasarkan senarai semak instrumen pemarkahan (Lampiran F). Seterusnya, draf-draf tersebut diterjemahkan ke dalam Bahasa Melayu dengan menggunakan kaedah *back translation*. Seramai enam orang guru bahasa dijemput untuk menjalankan tugas ini. Mereka merupakan guru Bahasa Melayu dan Bahasa Cina di sekolah rendah. Tiga orang guru bahasa menterjemahkan draf-draf tersebut daripada Bahasa Cina kepada Bahasa Melayu dan selepas itu tiga orang guru bahasa yang lain pula menterjemahkan versi Bahasa Melayu kepada Bahasa Cina semula. Pengkaji menyemak semula salinan asal dengan salinan Bahasa Cina yang baru dan melakukan sedikit pemurnian ke atas naskhah penterjemahan.

Untuk mengesahkan keberkesanan instrumen-instrumen pemarkahan yang dihasilkan, tiga orang pakar pentaksiran daripada badan pentaksiran dalam negara

diminta untuk menilai instrumen-instrumen pemarkahan tersebut dengan berasaskan pernyataan-pernyataan yang terdapat dalam borang pengesahan keberkesanan instrumen pemarkahan (Lampiran G). Borang tersebut dibentuk berpandukan manual pelaksanaan aktiviti pengesanan dan pencerapan daripada Lembaga Peperiksaan Malaysia (2002). Sebanyak tujuh pernyataan dikemukakan untuk menilai setiap instrumen pemarkahan. Senarai pernyataan tersebut ialah ketepatan konstruk yang diukur, kepentingan dari segi pembelajaran dan pengajaran, kebolehlaksanaan instrumen, kesesuaian instrumen pemarkahan dengan kebolehan murid Tahun Enam, kejelasan pembahagian tahap dan huraian kriteria, dan mutu keseluruhan instrumen pemarkahan tersebut. Pakar pentaksiran menilai instrumen pemarkahan karangan berkenaan berdasarkan urutan angka iaitu 4 (sangat memuaskan), 3 (memuaskan), 2 (tidak memuaskan) dan 1 (sangat tidak memuaskan). Keputusan penilaian yang diperoleh telah ditunjukkan dalam Jadual 3.5.

Dari segi peratusan persetujuan jitu (*percentage of exact agreement*) berdasarkan prosedur pemarkahan yang berlainan, peratusan persetujuan antara pakar adalah agak tinggi iaitu hampir kesemuanya melebihi 50% terutamanya persetujuan antara pakar A dengan pakar B (85.7%) (Lampiran H). Manakala peratusan persetujuan dalam satu poin (lihat Linn & Gronlund, 2000, p. 116) adalah 100% pada keseluruhannya bagi persetujuan antara pakar merentas instrumen pemarkahan yang berbeza. Peratusan persetujuan antara pakar yang agak tinggi ini menandakan para pakar memberi nilai angka yang jitu apabila menilai instrumen pemarkahan yang terlibat (Tinsley & Weiss, 2000). Secara kesimpulan, dapatan keputusan penilaian pakar menunjukkan bahawa mutu bagi kesemua instrumen pemarkahan karangan yang dihasilkan adalah memuaskan.

Jadual 3.5

Keputusan Penilaian Pakar Pentaksiran Untuk Empat Jenis Instrumen Pemarkahan

Pernyataan	Aspek Kandungan Dan Organisasi						Aspek Penggunaan Bahasa Dan Mekanis					
	Kaedah Holistik			Kaedah Analitik			Kaedah Holistik			Kaedah Analitik		
	Pakar			Pakar			Pakar			Pakar		
	A	B	C	A	B	C	A	B	C	A	B	C
Menepati konstruk yang diukur	4	4	4	4	4	4	4	4	4	4	4	4
Kepentingan pengajaran dan pembelajaran	4	4	4	4	4	4	4	4	4	4	4	4
Kebolehlaksanaan Instrumen	4	4	3	4	4	4	4	4	4	4	4	3
Kesesuaian dari segi aras kesukaran	4	4	4	4	4	3	4	4	3	4	4	3
Kejelasan dari segi pembahagian tahap	4	4	4	4	4	4	4	4	4	4	4	4
Kejelasan dari segi huraian kriteria	4	4	3	4	4	3	4	4	3	4	4	3
Kualiti keseluruhan instrumen	4	3	3	4	3	3	4	3	4	4	3	3
Jumlah	28	27	25	28	27	25	28	27	26	28	27	24

Nota. Penilaian pakar berdasarkan angka 1 hingga 4 iaitu 1 = sangat tidak memuaskan, 2 = tidak memuaskan, 3 = memuaskan dan 4 = sangat memuaskan.

Namun begitu, pakar pentaksiran juga mengemukakan beberapa cadangan untuk memantapkan lagi instrumen-instrumen berkenaan. Ini termasuklah memperbetulkan perkataan yang mungkin mengelirukan huraian kriteria iaitu aksara kepada karakter, markat ditukar kepada markah, dan menggunakan perenggan baru untuk memisahkan kriteria yang mempunyai lebih daripada satu ayat untuk memudahkan pemahaman pemeriksa. Sejajar dengan itu, pengkaji telah membuat pengubahsuaian dan pemurnian terhadap instrumen-instrumen berkenaan

berdasarkan cadangan-cadangan tersebut dalam Bahasa Cina dan Bahasa Melayu (lihat Lampiran I 1 hingga I 4 dan Lampiran I 1a hingga I 4a).

3.4.5 Kajian rintis

Tujuan kajian rintis dijalankan adalah untuk menilai kebolehlaksanaan semua instrumen yang dibina (Chua, 2006) dan mengenal pasti darjah kemunasabahan kajian sebenar. Selain mengenal pasti kebolehpercayaan dan kesahan skor tugasan karangan dan instrumen pemarkahan, aspek-aspek kekaburan, kekeliruan atau ketidakcukupan perwakilan instrumen karangan dan pemarkahan perlu dikesan agar membolehkan pemurnian yang sewajarnya dijalankan (Ary et al., 2006) sebelum kajian sebenar dilakukan.

Kajian rintis ini telah dijalankan di salah sebuah SJK(C) di negeri Perak memandangkan sekolah tersebut berdekatan dengan kawasan kajian sebenar. Selain itu, tahap pencapaian kemahiran menulis murid Tahun Enam di sekolah tersebut adalah pelbagai dan lebih kurang sama dengan sampel kajian di tempat kajian sebenar. Oleh itu, murid Tahun Enam sekolah tersebut dan sampel kajian boleh dianggap memiliki ciri-ciri populasi yang sama atau hampir sama.

Untuk mengenal pasti kebolehlaksanaan dan kecukupcakupan tugasan karangan, satu soal selidik telah dibentuk berpandukan manual pengesanan instrumen LPM (2002) bagi mengesan sejauh mana kekuatan dan kelemahan soalan tugasan karangan yang diuji (Lampiran J dan Lampiran J a). Soal selidik ini mengandungi 10 pernyataan atau item. Calon diminta memberi hanya satu respons untuk setiap item mengikut skala yang membawa maksud: 4 (sangat setuju), 3 (setuju), 2 (kurang setuju) dan 1(tidak setuju). Soal selidik ini telah dijawab oleh

calon sejurus selepas mereka tamat menulis tugasan karangan berkenaan. Dalam kajian ini, dua buah karangan berbeza bentuk telah digunakan untuk mentaksir kemahiran menulis calon iaitu bentuk pendedahan (topik sekolah) dan bentuk naratif (topik hadiah) (Lampiran K). Pengkaji telah memberikan sedikit penjelasan yang diperlukan sebelum calon menjawab soal selidik tersebut. Maklum balas calon dalam soal selidik ini meliputi tiga kriteria utama iaitu: (1) kesesuaian soalan karangan dari segi peruntukan masa dan tahap kesukaran, (2) kejelasan soalan karangan bagi keseluruhan item, dan (3) keakuran soalan karangan dari aspek keadilan dan peluang belajar.

Soal selidik ini mengesan perkara-perkara berikut yang diselitkan dalam 10 item seperti berikut:

1. Sejauh manakah tanggapan calon tentang kesesuaian peruntukan masa yang diberikan? Item yang digunakan ialah item no.1 dalam soal selidik calon.
2. Sejauh manakah tanggapan calon tentang aras kesukaran sama seperti yang dihajati? Item yang digunakan ialah item no.3 dalam soal selidik calon.
3. Sejauh manakah kesesuaian soalan karangan dari aspek kejelasan? Item yang digunakan ialah item no.2, 4, 5, 7 dan no.10 dalam soal selidik calon.
4. Sejauh manakah kesesuaian soalan karangan dari aspek keadilan dan peluang belajar? Item yang digunakan ialah item no.6, 8 dan no.9 dalam soal selidik calon.

Seramai 30 orang calon telah memberi respons kepada 10 pernyataan dalam soal selidik ini. Untuk memudahkan pentafsiran, pengiraan respons bagi skala sangat setuju dan setuju disatukan. Secara keseluruhan, dapatan tentang tajuk karangan yang berkaitan dengan kebersihan sekolah (berunsur pendedahan) agak memuaskan memandangkan 231 respons (77.0%) daripada calon bersetuju dengan pernyataan-pernyataan yang dikemukakan berbanding dengan 67 respons (22.3%) kurang setuju dan dua respons (0.7%) tidak setuju merentas kriteria kesesuaian, kejelasan dan keakuran (lihat Jadual 3.6).

Jika ditinjau daripada setiap kriteria utama, kriteria kejelasan mendapat respons setuju daripada calon yang paling tinggi iaitu 78.7% (118 respons) dan 21.3% (32 respons) menyatakan kurang bersetuju. Ini diikuti oleh kriteria kesesuaian iaitu 76.7% (46 respons) menyatakan setuju dan 20.0% (12 respons) kurang setuju di samping dua respons tidak setuju. Manakala bagi kriteria keakuran, 74.4% calon (67 respons) menyatakan setuju berbanding dengan 25.6% (23 respons) yang tidak setuju.

Jika dilihat dari segi respons terhadap pernyataan dalam setiap kriteria, pada keseluruhannya, respons calon terhadap pernyataan-pernyataan dalam kriteria kejelasan adalah memuaskan kecuali pernyataan tentang rangsangan soalan dan kehendak tajuk yang mempunyai respons kurang setuju sebanyak 10 respons (33.3%) dan 11 respons (36.7%) masing-masing. Bagi kriteria kesesuaian pula, pernyataan tentang tahap kesukaran mendapat respons kurang setuju agak tinggi iaitu 9 (30%) dan satu respons tidak setuju. Namun begitu, pernyataan pengalaman dalam kriteria

keakuran mendapat respons kurang setuju yang paling tinggi iaitu 14 respons (46.7%) .

Secara kesimpulan, dapatan soal selidik ini menunjukkan bahawa kelemahan soalan ini mungkin lebih memirip kepada pernyataan-pernyataan tentang tahap kesukaran, kehendak tajuk, rangsangan soalan dan terutamanya pengalaman. Untuk mengenal pasti kelemahan tersebut, pengkaji telah menganalisis secara terperinci selok-belok soalan karangan dan skrip jawapan calon bersama-sama dengan pakar pemeriksa. Persetujuan telah dicapai iaitu perkataan 'konkrit' yang digunakan dalam rangsangan soalan ini ditukarkan kepada 'sesuai' supaya sejajar dengan tahap kesukaran dan pengalaman kebanyakan murid Tahun Enam.

Jadual 3.6

Keputusan Kajian Soal Selidik Untuk Karangan Berunsur Pendedahan Berdasarkan Bilangan Respons

Pernyataan	No. Item	Sangat Setuju	Setuju	Kurang Setuju	Tidak Setuju
KRITERIA KESESUAIAN					
Peruntukan masa	1	12	14	3	1
Tahap kesukaran	3	4	16	9	1
KRITERIA KEJELASAN					
Kehendak tajuk	2	12	7	11	-
Bahasa	4	11	14	5	-
Arahan	5	15	12	3	-
Kehendak soalan	7	20	7	3	-
Rangsangan soalan	10	11	9	10	-
KRITERIA KEAKURAN					
Pengalaman	6	5	11	14	-
Pengetahuan diajar	8	12	14	4	-
Bentuk karangan diajar	9	9	16	5	-
Jumlah (%)	10	111(37.0)	120(40.0)	67(22.3)	2(0.7)

Nota. $n = 30$.

Secara keseluruhan, dapatan soal selidik tentang tajuk karangan yang berkaitan dengan sebuah hadiah (berunsur naratif) adalah lebih memuaskan berbanding karangan berunsur pendedahan memandangkan 84.0% calon (252 respons) bersetuju dengan pernyataan-pernyataan yang dikemukakan merentas kriteria kesesuaian, kejelasan dan keakuran. Respons yang selebihnya iaitu 15.0% (45 respons) menyatakan kurang setuju dan 1.0% (3 respons) menyatakan tidak setuju (lihat Jadual 3.7).

Jika ditinjau dari segi kriteria utama, 86.7% (130 respons) calon bersetuju dengan kriteria kejelasan berbanding dengan kriteria kesesuaian dan kriteria keakuran iaitu masing-masing 81.7% (49 respons) dan 81.1%, (73 respons). Peratusan respons bagi kurang setuju untuk kriteria kesesuaian dan kriteria keakuran adalah sama iaitu masing-masing 16.7% manakala bagi kriteria kejelasan ialah 13.3%. Selain itu, kriteria keakuran juga mendapat dua respons bagi tidak setuju sementara kriteria kesesuaian mendapat satu respons.

Pada keseluruhannya, respons calon terhadap pernyataan-pernyataan dalam ketiga-tiga kriteria utama adalah memuaskan kecuali pernyataan mengenai tahap kesukaran, rangsangan soalan dan pengetahuan diajar. Pernyataan tentang pengetahuan diajar mendapatkan respons kurang setuju yang paling tinggi iaitu 30% (9 respons) dan satu respons tidak setuju.

Secara kesimpulan, dapatan keputusan soal selidik menunjukkan bahawa kelemahan soalan ini mungkin berkaitan dengan kriteria keakuran terutamanya pernyataan tentang pengetahuan diajar. Di samping itu, perhatian juga perlu diberikan kepada tahap kesukaran dan rangsangan soalan. Setelah menganalisis skrip jawapan calon, didapati walaupun calon mampu melengkapkan karangan tetapi rata-rata tidak dapat menggambarkan 'hadiah di luar dugaan' dengan memuaskan. Hasil perbincangan dengan pakar pemeriksa, pengkaji mengubahsuaikan tajuk karangan daripada 'sebuah hadiah di luar dugaan' kepada 'sebuah hadiah yang istimewa'.

Jadual 3.7

Keputusan Kajian Soal Selidik Untuk Karangan Berunsur Naratif Berdasarkan Bilangan Respons

Pernyataan	No. Item	Sangat Setuju	Setuju	Kurang Setuju	Tidak Setuju
KRITERIA KESESUAIAN					
Peruntukan masa	1	17	9	3	1
Tahap kesukaran	3	7	16	7	-
KRITERIA KEJELASAN					
Kehendak tajuk	2	10	16	4	-
Bahasa	4	15	14	1	-
Arahan	5	17	10	3	-
Kehendak soalan	7	11	14	5	-
Rangsangan soalan	10	8	15	7	-
KRITERIA KEAKURAN					
Pengalaman	6	9	16	4	1
Pengetahuan diajar	8	4	16	9	1
Bentuk karangan diajar	9	7	21	2	-
Jumlah (%)	10	105(35.0)	147(49.0)	45(15.0)	3(1.0)

Nota. $n = 30$.

Untuk pentaksiran yang menggunakan instrumen pemarkahan atau penilaian subjektif, maklumat mengenai kebolehpercayaan antara pemeriksa bagi satu set

pemarkahan adalah penting dalam menentukan kesahan dan kebolehan generalisasi sesuatu keputusan kajian (Tinsley & Weiss, 2000). Dalam kajian ini, sekumpulan pakar pemeriksa yang berpengalaman menilai karangan Bahasa Cina sekolah rendah telah diminta untuk menilai kedua-dua buah tugasan karangan. Mereka memeriksa skrip jawapan karangan bagi 30 orang murid dengan instrumen pemarkahan yang berlainan iaitu berdasarkan aspek dan kaedah pemarkahan yang berlainan. Setiap aspek dan kaedah pemarkahan bagi karangan berlainan tugasan dinilai oleh dua orang pakar pemeriksa dalam satu kumpulan. Keputusan bagi kebolehpercayaan antara pemeriksa berdasarkan prosedur pemarkahan yang berlainan telah ditunjukkan dalam Jadual 3.8.

Jadual 3.8

Min, Sisihan Piawai Dan Kebolehpercayaan Antara Pemeriksa Untuk Kaedah Dan Aspek Pemarkahan Yang Berlainan Berdasarkan Karangan Yang Berlainan

	Karangan Berunsur Pendedahan*			Karangan Berunsur Naratif*		
Prosedur Pemarkahan	*M*	*SD*	*r*	*M*	*SD*	*r*
Kaedah Holistik						
Aspek Kandungan dan Organisasi	5.10 5.10	1.86 1.77	.921**	5.20 5.20	1.82 1.70	.901**
Aspek Penggunaan Bahasa dan Mekanis	5.50 5.60	1.73 1.73	.914**	5.65 5.45	1.72 1.54	.915**
Kaedah Analitik						
Aspek Kandungan dan Organisasi	6.40 6.30	2.06 2.08	.952**	5.65 6.00	1.72 1.78	.944**
Aspek Penggunaan Bahasa dan Mekanis	6.65 6.35	2.21 2.11	.943**	6.30 5.75	1.59 1.65	.931**

Nota. **n*= 30. *r* = pekali korelasi *Pearson*.
***p* < .01.

Secara keseluruhan, keputusan pekali kebolehpercayaan (*r*) bagi kedua-dua tugasan karangan adalah berada dalam lingkungan yang boleh diterima malah memuaskan dari segi penilaian respons terbuka yang dilakukan oleh pemeriksa

manusia. Berdasarkan prosedur pemarkahan yang berlainan, anggaran kebolehpercayaan antara pemeriksa bagi karangan bentuk pendedahan mempunyai julat antara .91 hingga .95 manakala bagi bentuk naratif adalah antara .90 hingga .94 dengan aras signifikan $p < .01$ masing-masing. Ini menunjukkan bahawa anggaran pekali kebolehpercayaan bagi kedua-dua tugasan karangan adalah tidak jauh berbeza antara satu sama lain.

Jika dilihat daripada tugasan karangan yang berlainan berdasarkan kaedah pemarkahan, nilai pekali kebolehpercayaan bagi kaedah analitik yang ditaksir dengan aspek pemarkahan yang berlainan adalah lebih tinggi sedikit (.931, .943, .944 dan .952) daripada kaedah holistik yang ditaksir dengan aspek pemarkahan yang sama (.901, .914, .915 dan .921). Sementara itu, bagi karangan berlainan tugasan berdasarkan aspek pemarkahan pula, nilai pekali kebolehpercayaan yang dapat dilihat adalah bercampur-campur dengan aspek kandungan dan organisasi serta kaedah analitik mempunyai nilai paling tinggi (.944 dan .952). Pada asasnya, nilai kebolehpercayaan antara pemeriksa yang tinggi boleh dianggap bahawa hubungan antara objek yang dinilai adalah sama merentas pemeriksa (Tinsley & Weiss, 2000). Dapatan keputusan ini juga memperlihatkan bahawa pemeriksa tidak mengalami kesukaran dalam melakukan penilaian yang boleh dipercayai tentang ciri-ciri yang terdapat dalam kedua-dua tugasan karangan tersebut berdasarkan instrumen pemarkahan yang berlainan.

Untuk menentukan bukti takat kesahan tugasan karangan yang digunakan berkaitan dengan gagasan kemahiran menulis, skor bagi aspek pemarkahan yang berlainan berdasarkan kaedah holistik dan analitik telah dikorelasikan dengan skor karangan bahasa Cina PKSR semester pertama tahun 2008 dan skor penilaian guru

terhadap penguasaan kemahiran menulis murid (Lampiran L). Kriteria-kriteria penilaian guru adalah berdasarkan aspek-aspek penilaian karangan Bahasa Cina KBSR tahap II (KPM, 1997). Untuk kaedah holistik, julat pekali korelasi antara aspek pemarkahan dengan PKSR (Jadual 3.9) yang didapati adalah antara .77 hingga .87. Manakala kolerasi antara aspek pemarkahan dengan penilaian guru adalah antara .85 hingga .91. Untuk kaedah analitik pula, keputusannya adalah antara .82 hingga .90, sementara julat pekali korelasi adalah antara .77 hingga .89 bagi korelasi aspek pemarkahan dan penilaian guru (lihat Jadual 3.10). Pada keseluruhannya, boleh dikatakan hampir kesemua pekali korelasi berada di atas .80 dengan aras signifikan $p < .01$. Korelasi yang agak tinggi ini boleh diandaikan bahawa mereka mengukur konstruk yang didasari yang sama (Mohd. Najib Ghafar, 1997; Mokhtar Ismail, 1995; Kubiszyn & Borich, 2003).

Jadual 3.9

Matriks Korelasi Antara Aspek Pemarkahan Berdasarkan Kaedah Holistik Dengan Keputusan PKSR Dan Penilaian Guru

	Variabel	2	3	4	5	6	7	8
1.	P. kan/org	.919**	.981**	.800**	.794**	.819**	.774**	.866**
2.	P. bhs/mek		.978**	.710**	.822**	.784**	.813**	.854**
3.	P. seluruh			.773**	.824**	.819**	.809**	.878**
4.	N. kan/org				.897**	.976**	.826**	.878**
5.	N. bhs/mek					.972**	.862**	.888**
6.	N. seluruh						.866**	.907**
7.	PKSR							.856**
8.	Guru							

Nota. $n = 30$. P. kan/org = karangan bentuk pendedahan serta aspek kandungan dan organisasi; P. bhs/mek = karangan bentuk pendedahan serta aspek penggunaan bahasa dan mekanis; P. seluruh = seluruh karangan bentuk pendedahan; N. kan/org = karangan bentuk naratif serta aspek kandungan dan organisasi; N. bhs/mek = karangan bentuk naratif serta aspek penggunaan bahasa dan mekanis; N.

seluruh = seluruh karangan bentuk naratif; PKSR = Penilaian Kemajuan Sekolah Rendah; Guru = penilaian guru.
**$p < .01$.

Untuk menentukan takat kesahan skor instrumen pemarkahan kaedah holistik dan analitik yang dibina dalam kajian ini (instrumen kajian), analisis korelasi telah dijalankan antara instrumen kajian dan instrumen pemarkahan holistik dan analitik yang digunakan oleh badan peperiksaan (instrumen badan peperiksaan). Untuk

Jadual 3.10

Matriks Korelasi Antara Aspek Pemarkahan Berdasarkan Kaedah Analitik Dengan Keputusan PKSR Dan Penilaian Guru

	Variabel	2	3	4	5	6	7	8
1.	P. kan/org	.846**	.959**	.704**	.724**	.722**	.820**	.809**
2.	P. bhs/mek		.962**	.727**	.745**	.750**	.817**	.773**
3.	P. seluruh			.745**	.765**	.766**	.852**	.823**
4.	N. kan/org				.901**	.979**	.894**	.809**
5.	N. bhs/mek					.969**	.853**	.889**
6.	N. seluruh						.901**	.864**
7.	PKSR							.849**
8.	Guru							

Nota. $n = 30$. P. kan/org = karangan bentuk pendedahan serta aspek kandungan dan organisasi; P. bhs/mek = karangan bentuk pendedahan serta aspek penggunaan bahasa dan mekanis; P. seluruh = seluruh karangan bentuk pendedahan; N. kan/org = karangan bentuk naratif serta aspek kandungan dan organisasi; N. bhs/mek = karangan bentuk naratif serta aspek penggunaan bahasa dan mekanis; N. seluruh = seluruh karangan bentuk naratif; PKSR = Penilaian Kemajuan Sekolah Rendah; Guru = penilaian guru.
**$p < .01$.

tujuan tersebut, skor pemarkahan holistik bagi aspek kandungan dan organisasi serta aspek penggunaan bahasa dan mekanis untuk instrumen kajian telah dijumlahkan agar satu perbandingan yang menyeluruh boleh dibuat. Begitu juga skor pemarkahan analitik untuk instrumen kajian. Manakala instrumen badan peperiksaan yang digunakan merupakan instrumen pemarkahan holistik dan analitik yang digunakan

untuk mentaksir karangan Bahasa Cina peringkat UPSR. Pekali korelasi yang tinggi antara kedua-dua instrumen tersebut dianggap bukti tentang kesahan berasaskan kriteria. Analisis korelasi mendapati pekali korelasi instrumen kajian dan instrumen badan peperiksaan adalah tinggi (Jadual 3.11). Untuk pemarkahan holistik, pekali korelasi bagi karangan berunsur pendedahan dan naratif masing-masing ialah .83 ($p < .01$) dan .81 ($p < .01$). Sementara itu, pekali korelasi berdasarkan pemarkahan analitik bagi karangan berunsur pendedahan dan naratif ialah .93 ($p < .01$) dan .86 ($p < .01$).

Jadual 3.11

Korelasi Skor Instrumen Kajian Dan Instrumen Badan Peperiksaan Berdasarkan Pemarkahan Holistik Dan Analitik Untuk Tugasan Karangan Yang Berlainan

Kaedah Pemarkahan	Pendedahan*	Naratif*
PEMARKAHAN HOLISTIK		
Instrumen kajian Instrumen badan peperiksaan	.833**	.812**
PEMARKAHAN ANALITIK		
Instrumen kajian Instrumen badan peperiksaan	.929**	.864**

Nota. *n= 30.
**$p < .01$.

3.4.6 Pemeriksa

Seramai 12 orang pemeriksa yang terdiri daripada guru SJK(C) telah melibatkan diri dalam kerja pemarkahan skrip karangan sampel kajian (Jadual 3.12). Lapan daripada mereka adalah perempuan (66.6%) dan selainnya lelaki. Tempoh pengalaman mengajar mata pelajaran Bahasa Cina KBSR mereka adalah antara 13

hingga 30 tahun (*M* = 21.08, *SD* = 5.14). Mereka berpengalaman memeriksa karangan Bahasa Cina di sekolah dan juga merupakan pemeriksa kertas karangan Bahasa Cina UPSR. Julat umur mereka adalah antara 38 hingga 55 tahun (*M* = 46.08, *SD* = 5.60).

Jadual 3.12

Butir-Butir Mengenai Pemeriksa

Bil	Pemeriksa	Jantina	Umur	Pengalaman Mengajar (Tahun)
1.	r1	P	44	19
2.	r2	P	46	22
3.	r3	L	55	30
4.	r4	L	38	13
5.	r5	P	48	20
6.	r6	P	41	17
7.	r7	P	55	30
8.	r8	P	46	20
9.	r9	P	50	25
10.	r10	L	40	16
11.	r11	L	49	20
12.	r12	P	41	21

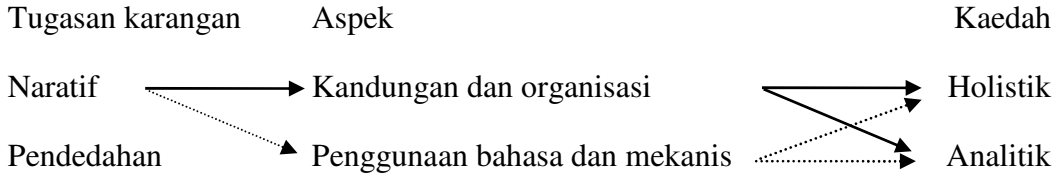
Nota. r1= pemeriksa pertama, r2= pemeriksa kedua dan seterusnya ; P = perempuan; L= lelaki.

3.4.7 Sampel-sampel karangan

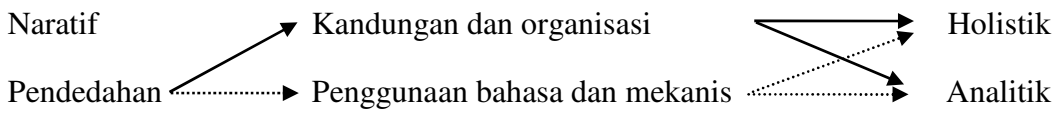
Untuk sesi latihan pemeriksa, pengkaji telah menyediakan 8 set sampel karangan mengikut tugasan karangan, aspek dan kaedah pemarkahan masing-masing. Tujuannya adalah untuk menentukan ketekalan pemarkahan antara pemeriksa serta pemberian markah yang tepat dan adil. Setiap set mempunyai 10 buah karangan yang dipilih khas dan dikenal pasti meliputi tiga tahap pencapaian yakni tinggi, sederhana dan rendah berserta kertas *anchor* bagi setiap tahap pencapaian. Tugasan karangan tersebut adalah dipilih daripada tugasan karangan murid Tahun Enam yang tidak

mengambil bahagian dalam kajian ini (lihat Saddle & Graham, 2005). Berikut adalah senarai 8 set sampel karangan yang disediakan berdasarkan Rajah 3.4 dan Rajah 3.5:

1. karangan berunsur naratif, aspek kandungan dan organisasi serta kaedah holistik,
2. karangan berunsur naratif, aspek kandungan dan organisasi serta kaedah analitik,
3. karangan berunsur naratif, aspek penggunaan bahasa dan mekanis serta kaedah holistik,
4. karangan berunsur naratif, aspek penggunaan bahasa dan mekanis serta kaedah analitik,
5. karangan berunsur pendedahan, aspek kandungan dan organisasi serta kaedah holistik,
6. karangan berunsur pendedahan, aspek kandungan dan organisasi serta kaedah analitik,
7. karangan berunsur pendedahan, aspek penggunaan bahasa dan mekanis serta kaedah holistik, dan
8. karangan berunsur pendedahan, aspek penggunaan bahasa dan mekanis serta kaedah analitik.



Rajah 3.4. Karangan berunsur naratif dimarkah dengan kaedah dan aspek pemarkahan yang berlainan.



Rajah 3.5. Karangan berunsur pendedahan dimarkah dengan kaedah dan aspek pemarkahan yang berlainan.

3.4.8 Pemarkahan sampel karangan

Kerja pemarkahan ini diadakan sebanyak 2 kali pada hujung minggu yang keseluruhannya memakan masa 4 hari (lihat Lampiran M 1 dan Lampiran M 2). Pada kali pertama, pemarkahan skrip jawapan adalah berkaitan dengan prosedur pemarkahan iaitu aspek kandungan dan organisasi serta kaedah holistik dan analitik berdasarkan karangan berlainan tugasan. Manakala pada kali kedua, pemarkahan skrip jawapan pula adalah mengenai aspek penggunaan bahasa dan mekanis serta kaedah analitik dan holistik berdasarkan karangan berlainan tugasan. Untuk kerja pemarkahan kaedah analitik, kedua-dua aspek kandungan dan organisasi perlu ditaksir secara berasingan, kemudian kedua-dua bahagian markah tersebut dijumlahkan. Begitu juga dengan aspek penggunaan bahasa dan mekanis, kedua-duanya adalah ditaksir secara berasingan dan kemudian dijumlahkan. Oleh itu, kerja penilaian dan pengiraan untuk kaedah analitik adalah lebih sedikit berbanding dengan kaedah holistik.

Dalam pemeriksaan skrip kali pertama, pengkaji akan memberikan taklimat ringkas mengenai tujuan pemarkahan, prosedur kerja dan sasaran kerja kepada para pemeriksa. Kemudian, program latihan diteruskan dengan sesi pendedahan tentang prosedur-prosedur pemarkahan untuk memperkemas dan mengukuhkan lagi kemahiran pemarkahan pemeriksa bagi aspek kandungan dan organisasi berdasarkan kaedah holistik dan analitik. Walaupun pemeriksa hanya akan menggunakan salah

satu kaedah untuk kerja pemarkahan, tetapi pendedahan sebegini akan memberi gambaran yang menyeluruh tentang perbezaan antara kedua-dua kaedah tersebut.

Berdasarkan reka bentuk eksperimen teori G dalam kajian ini, pemeriksa adalah tersarang dalam tugasan karangan yang berlainan. Oleh itu, sebelum memulakan sesi latihan awal, pemeriksa diagihkan ke dalam kumpulan kecil mengikut tugasan karangan yang berlainan. Dua belas orang pemeriksa yang hadir dibahagikan secara rawak ke dalam empat kumpulan yang ditetapkan. Pada sesi latihan awal, setiap kumpulan akan diberikan satu set sampel karangan yang mengandungi 10 buah karangan mengikut jadual pengagihan kerja. Misalnya, kumpulan A akan memeriksa sampel karangan berunsur naratif untuk aspek kandungan dan organisasi serta kaedah holistik. Masa yang diperuntukkan untuk menyiapkan pemarkahan ialah 30 minit. Setelah itu, sesi perbincangan diadakan untuk memperjelas prosedur pemarkahan dan penyelarasan markah. Bagi kumpulan yang menjalankan pemarkahan dengan kaedah analitik, pemeriksa perlu memberi markah secara berasingan untuk aspek kandungan dan aspek organisasi, kemudian menjumlahkan kedua-dua bahagian markah.

Selepas sesi latihan awal, pemeriksa menilai 10 buah lagi tugasan karangan yang sama untuk tujuan pengukuhan. Masa pemarkahan yang diberikan adalah sama iaitu 30 minit. Setelah selesai pemarkahan, ketiga-tiga pemeriksa dalam setiap kumpulan membincang dan menyelaraskan markah antara satu sama lain. Setelah tamat latihan sesi kedua, setiap pemeriksa diberikan 120 skrip karangan mengikut tugasan yang diagihkan. Mereka diberi masa lebih kurang 4 jam untuk menyiapkan kerja pemeriksaan. Markah yang diberi perlu dicatat di sebelah atas sudut kanan

skrip karangan dan dipindahkan ke dalam borang pemarkahan yang disediakan (Lampiran N 1 hingga Lampiran N 6). Sesi penyemakan semula markah dengan memadankan markah dalam skrip karangan dan borang pemarkahan dijalankan secara berpasangan untuk mengesahkan ketepatan pengisian markah. Jadual 3.13 menunjukkan pengagihan tugas bagi setiap kumpulan untuk kerja pemarkahan karangan kali pertama. Manakala Jadual 3.14 menunjukkan pengagihan tugas pemeriksaan bagi setiap kumpulan untuk kerja pemarkahan kali kedua. Jumlah skrip karangan yang dapat dikumpul bagi setiap sesi daripada setiap kumpulan ialah 360 dan jumlah keseluruhan skrip yang dapat dikumpul ialah 1440.

Langkah-langkah latihan pemeriksa dan prosedur kerja untuk pemeriksaan skrip kali kedua adalah sama seperti dalam kali pertama. Namun, pada kali ini setiap kumpulan akan mendapat pengagihan tugasan yang berlainan dalam erti kata bahawa setiap kumpulan akan memarkah dengan tugasan karangan yang berlainan dengan menggunakan aspek dan kaedah pemarkahan yang berlainan. Ini bermakna pada

Jadual 3.13

Pengagihan Kerja Pemeriksaan Skrip Pada Kali Pertama

Kumpulan	A	B	C	D
Tugasan karangan	Naratif	Pendedahan	Naratif	Pendedahan
Kaedah pemarkahan	Holistik	Holistik	Analitik	Analitik
Aspek pemarkahan	Kandungan dan Organisasi	Kandungan dan Organisasi	Kandungan dan Organisasi	Kandungan dan Organisasi
Jumlah skrip	120	120	120	120
Pemeriksa	P1, P2, P3	P4, P5, P6	P7, P8, P9	P10, P11, P12

Data dikumpul	360	360	360	360

Jadual 3.14

Pengagihan Kerja Pemeriksaan Skrip Pada Kali Kedua

Kumpulan	A	B	C	D
Tugasan karangan	Pendedahan	Naratif	Pendedahan	Naratif
Kaedah pemarkahan	Analitik	Analitik	Holistik	Holistik
Aspek pemarkahan	Penggunaan Bahasa dan Mekanis	Penggunaan Bahasa dan Mekanis	Penggunaan Bahasa dan Mekanis	Penggunaan Bahasa dan Mekanis
Jumlah skrip	120	120	120	120
Pemeriksa	P1, P2, P3	P4, P5, P6	P7, P8, P9	P10, P11, P12
Data dikumpul	360	360	360	360

pemarkahan kali kedua ini, pemeriksa kumpulan A akan menilai sampel karangan berunsur pendedahan dengan menggunakan aspek penggunaan bahasa dan mekanis berdasarkan kaedah analitik. Begitu juga dengan kumpulan-kumpulan lain yang akan menggunakan prosedur pemarkahan dan tugasan karangan yang berlainan dalam pemarkahan.

3.5 Prosedur Kajian Teori G

Proses asas dalam analisis teori G tentang sesuatu persoalan melibatkan dua peringkat iaitu kajian G dan kajian D. Dalam satu ujian yang sama, reka bentuk eksperimen bagi kedua-dua kajian tersebut boleh jadi sama atau tidak sama berdasarkan tujuan kajian. Dalam kajian G, tugas utama pengkaji adalah untuk mengenal pasti objek pengukuran. Objek pengukuran dalam kajian ini ialah

kebolehan menulis calon. Seterusnya, pengkaji mengenal pasti dan menetapkan faktor-faktor yang mempengaruhi skor ujian sebagai faset pengukuran. Kajian ini telah menggariskan empat faset pengukuran iaitu tugasan karangan (dua paras), pemeriksa (tiga paras), aspek pemarkahan (dua paras) dan kaedah pemarkahan (dua paras). Oleh itu, setiap calon akan memperoleh 24 skor pada keseluruhannya.

Reka bentuk eksperimen dalam kajian ini agak kompleks kerana pengkaji perlu meninjau kesan tugasan karangan dan pemeriksa berdasarkan dua aspek pemarkahan dan dinilai pula dengan dua kaedah pemarkahan yang berlainan. Kajian ini menggunakan konsep univariat dalam teori G untuk membuat anggaran. Oleh itu, pada hakikatnya pengkaji telah melakukan empat kali pengukuran melalui kajian G terhadap kesan tugasan karangan dan pemeriksa.

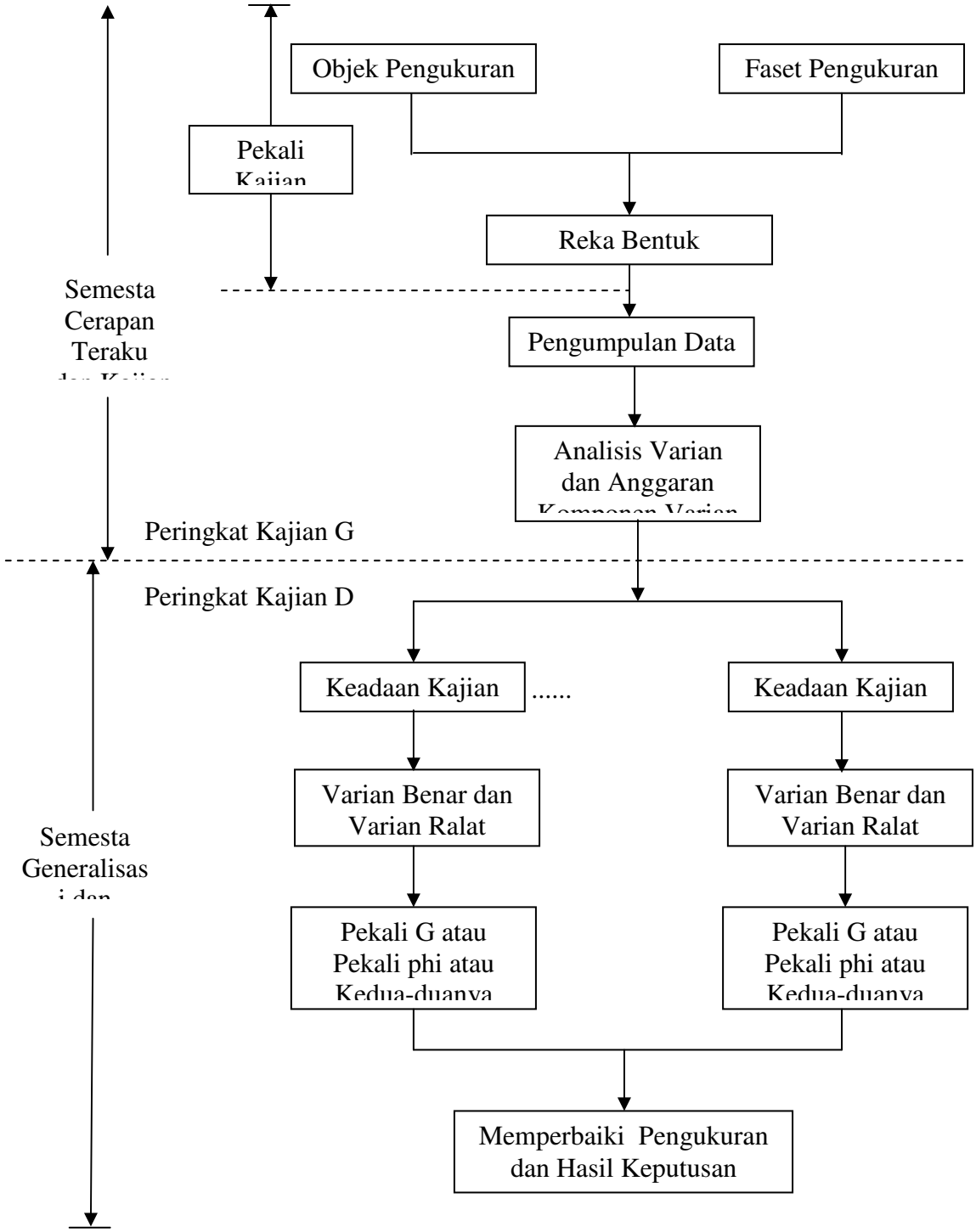
Selepas mengenal pasti objek dan faset pengukuran, pengkaji perlu mengenal pasti semesta cerapan teraku (*universe of admissible observations*) oleh pengukuran iaitu cerapan-cerapan yang mana pembuat keputusan sudi menganggapnya sebagai sesuatu yang boleh saling bertukaran untuk tujuan membuat sesuatu keputusan (Shavelson & Webb, 1991). Seterusnya, pengkaji memilih reka bentuk eksperimen yang sesuai bagi pengukuran. Dalam teori G terdapat tiga jenis pendekatan reka bentuk eksperimen iaitu reka bentuk tersilang, reka bentuk tersarang dan reka bentuk separa tersarang. Dalam keadaan reka bentuk tersilang, objek pengukuran dapat diukur oleh semua paras yang dikenal pasti bagi semua faset pengukuran. Manakala dalam reka bentuk tersarang, objek pengukuran hanya dapat diukur oleh sebahagian daripada paras bagi satu atau lebih daripada satu faset pengukuran (Suen, 1990). Reka bentuk separa tersarang pula mengandungi kedua-dua kesan reka bentuk

tersilang dan tersarang (Shavelson & Webb, 1991, p. 52). Kajian ini menggunakan reka bentuk separa tersarang $p \times (r{:}t)$ iaitu pemeriksa tersarang dalam tugasan karangan berlainan bentuk mengikut kumpulan yang berbeza dan setiap pemeriksa akan menilai semua karangan calon dalam kumpulan tersebut. Manakala pemeriksa dan tugasan karangan adalah tersilang dengan calon.

Data yang dikumpul dalam kajian diproses dengan menggunakan program GENOVA (*GENeralized analysis Of Variance system*) dan ujian *t* serta statistik deskriptif seperti min dan sisihan piawai. Program GENOVA direka khas oleh Crick dan Brennan (1983) untuk menangani kelas yang besar bagi reka bentuk ANOVA yang seimbang dan lengkap antaranya seperti reka bentuk tersilang, tersarang dan separa tersarang. Pada asasnya, tujuan operasi kajian G adalah untuk mendapatkan anggaran kadar komponen varian daripada sumbernya iaitu objek pengukuran, faset pengukuran dan pelbagai kesan interaksi lain antara mereka dan seterusnya menyediakan keadaan generalisasi bagi kajian D. Secara ringkas, proses-proses kajian mengenai teori G digambarkan dalam Rajah 3.6.

Menurut Brennan (2001, p. 9), kajian G dapat menganggar dan mentafsir komponen-komponen varian untuk merancang prosedur pengukuran secara terperinci dan seterusnya menghasilkan pekali G dan pekali phi (indeks kebergantungan) yang berbeza untuk tujuan eksperimen yang berlainan. Untuk menjalankan kajian D, spesifikasi semesta generalisasi perlu dirangka terlebih dahulu. Skop atau ruang lingkup semesta generalisasi adalah bergantung kepada bilangan faset pengukuran dan bilangan paras dalam setiap faset pengukuran. Keputusan generalisasi adalah dianggap kurang bermakna jika skop sesuatu semesta itu terlalu kecil. Dalam

peringkat ini, kajian D menyediakan dua jenis indeks kebolehpercayaan iaitu ralat relatif dan ralat mutlak yang masing-masing menghasilkan pekali G dan pekali phi. Pekali G sesuai digunakan untuk menganggar ujian rujukan norma manakala pekali phi pula sesuai untuk menghitung ujian rujukan kriteria. Pengendalian instrumen kajian yang disediakan dalam kajian ini adalah berdasarkan pendekatan ujian rujukan norma. Oleh itu, kajian ini akan membekalkan pekali G sahaja.



Rajah 3.6. Proses-proses kajian tentang teori G.

3.6 Langkah-Langkah Pengiraan Dalam Teori G

Langkah-langkah pengiraan untuk reka bentuk separa tersarang $p \times (r: t)$ dalam kajian ini meliputi kajian G dan kajian D.

3.6.1 Reka bentuk separa tersarang $p \times (r: t)$ model rawak dalam kajian G

Objek pengukuran dalam kajian ini ialah kebolehan menulis calon yang dilambangkan dengan *p* (*person*) dan dua faset pengukuran iaitu faset pemeriksa *r* (*rater*) dan tugasan karangan *t* (*task*). Semesta cerapan kajian ini terdiri daripada empat faset iaitu karangan berlainan tugasan, pemeriksa serta aspek dan kaedah pemarkahan. Kesan tugasan karangan dan pemeriksa akan dianggar melalui aspek dan kaedah pemarkahan yang berlainan. Pemeriksa tersarang dalam tugasan karangan yang berlainan bentuk mengikut kumpulan, dan setiap pemeriksa akan menilai semua tugasan karangan calon. Calon, tugasan karangan dan pemeriksa adalah dipilih secara rawak daripada semesta yang tak terhingga saiznya. Model pengukuran dalam teori G ini dikenali sebagai reka bentuk separa tersarang $p \times (r:t)$ model kesan rawak.

3.6.1.1 Model matematik dan andaian yang berkaitan

Dalam kajian ini, skor cerapan calon diwakili oleh X_{prt} iaitu skor calon *p* yang diberi oleh mana-mana pemeriksa tersarang dalam tugasan karangan berlainan bentuk dalam setiap kumpulan. Nilai jangkaan bagi variabel rawak, X_{prt} merentas tugasan dan pemeriksa berkaitan dengan proses di mana sesuatu tugasan dan

pemeriksa dipilih secara rawak daripada semesta (Brennan, 2001, p. 64) boleh ditunjukkan dengan:

$$\mu_p \equiv \underset{r}{E}\,\underset{t}{E}\, X_{prt} \tag{3.1}$$

Dalam persamaan (3.1), *E* ialah nilai jangkaan dan subskript *r* dan *t* masing-masing menandakan faset di mana jangkaan diambil. *E* digunakan untuk mewakili min jangka panjang (*long-run average*) bagi variabel rawak, X_{prt}. Manakala μ_p merupakan min skor calon yang diberi oleh mana-mana pemeriksa yang tersarang dalam tugasan karangan berlainan bentuk dalam setiap kumpulan. Ia dikenali sebagai skor semesta yang menyerupai skor benar dalam teori ujian klasik. Dengan cara yang sama, min seluruh bagi populasi dan semesta ialah:

$$\mu \equiv \underset{p}{E}\,\underset{r}{E}\,\underset{t}{E}\, X_{prt} \tag{3.2}$$

Min populasi bagi tugasan ialah:

$$\mu_t \equiv \underset{p}{E}\,\underset{r}{E}\, X_{prt} \tag{3.3}$$

dan min populasi bagi pemeriksa tersarang dalam tugasan ialah:

$$\mu_{rt} \equiv \underset{p}{E}\, X_{prt} \tag{3.4}$$

$$= \underset{p}{E}(\mu + v_p + v_t + v_{r:t} + v_{pt} + v_{pr:t})$$

$$= \mu + \underset{p}{E}\, v_p + v_t + v_{r:t} + \underset{p}{E}\, v_{pt} + \underset{p}{E}\, v_{pr:t}$$

$$= \mu + v_t + v_{r:t} \quad \text{di mana } \boldsymbol{E}\,(v_\alpha) = 0 \text{ (Brennan, 2001: 65)}$$

Manakala min semesta dan populasi bagi calon dan tugasan ialah:

$$\mu_{pt} \equiv \underset{r}{E}\, X_{prt} \tag{3.5}$$

Secara teori, jelas menunjukkan skor cerapan X_{prt} boleh dinyatakan dengan cara menjumlahkan lima parameter yang terlibat iaitu min skor μ, μ_p, μ_t, μ_{rt} dan μ_{pt}. Namun, semua parameter ini tidak boleh dicerap kerana adalah tidak mungkin bagi

setiap calon untuk menjawab semua tugasan karangan dalam semesta berkenaan ataupun semua calon dalam populasi memberi respons kepada tugasan berkenaan (Shavelson & Webb, 1991). Walaupun begitu, konsep tersebut dapat membantu membentuk model matematik yang biasa untuk reka bentuk teori G.

Reka bentuk kajian ini iaitu $p \times (r: t)$ boleh dinyatakan melalui pengungkaian skor cerapan dan kesan skor calon secara metematik. Kedua-dua pernyataan ini bersifat tautologi yakni mereka hanya merupakan perbezaan cara dalam menyatakan pengungkaian yang sama (Brennan, 1983 & 2001). Peleraian skor cerapan dan kesan skor calon untuk reka bentuk separa tersarang $p \times (r:t)$ model linear (Brennan, 1983, p. 28) boleh ditunjukkan seperti berikut:

Skor Cerapan	Kesan Skor
$X_{prt} = \mu$	(μ= min seluruh bagi skor semesta calon)
$+ (\mu_p - \mu)$	(v_p= kesan calon)
$+ (\mu_t - \mu)$	(v_t= kesan bentuk tugasan)
$+ (\mu_{rt} - \mu_t)$	($v_{r:t}$= kesan pemeriksa tersarang dalam tugasan)
$+ (\mu_{pt} - \mu_p - \mu_t + \mu)$	(v_{pt} = kesan interaksi calon dan tugasan)
$+ (X_{prt} - \mu_{pt} - \mu_{rt} + \mu_t)$	($v_{pr:t}$= kesan reja)

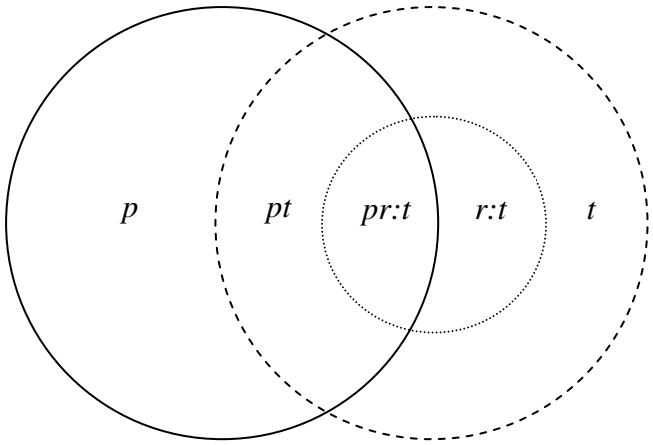
atau

$$X_{prt} = \mu + \mu_p + \mu_t + \mu_{r:t} + \mu_{pt} + \mu_{pr:t} \quad (3.6)$$

$$= \mu + v_p + v_t + v_{r:t} + v_{pt} + v_{pr:t} \quad (3.7)$$

Menurut Brennan (2001, p. 54), kesan sesuatu reka bentuk boleh dibahagikan kepada kesan utama dan kesan interaksi. Untuk reka bentuk $p \times (r: t)$ dalam kajian ini, kesan utama terdiri daripada v_p , v_t dan $v_{r:t}$ sementara v_{pt} dan $v_{pr:t}$ merupakan

kesan interaksi (kecuali μ). Dalam teori G, kesan tersarang seperti *r:t* dianggap sebagai kesan utama kerana menghubungkaitkan setiap faset dengan kesan utama amat bermakna dalam pendekatan teori G (Brennan, 2001). Cronbach et al. (1972) dan Shavelson dan Webb (1991) melambangkan kesan interaksi tiga hala reka bentuk separa tersarang sebagai $v_{pr,prt,e}$. Manakala Brennan (2001: 54) menganggap ralat reja (*residual error*) telah berbaur sama sekali dengan interaksi berkenaan, oleh itu adalah memadai mewakilinya dengan $v_{pr:t}$. Begitu juga kesan interaksi dua hala dilambangkan dengan $v_{r:t}$ daripada menggunakan $v_{r,\,rt}$. Tesis ini pada keseluruhannya akan menggunakan lambang Brennan sebagai asas rujukan. Secara mudah, simbol *p*,



Rajah 3.7. Gambar rajah Venn menunjukkan pengungkaian sumber-sumber keberubahan untuk reka bentuk separa tersarang dua faset $p \times (r:t)$ kesan rawak.

t dan *r:t* dianggap sebagai mewakili kesan utama, manakala *pt* dan *pr:t* menandakan kesan interaksi. Sumber-sumber kesan keberubahan berkenaan sesuai digambarkan melalui gambar rajah Venn (Brennan, 2001). Bagaimanapun, magnitud kawasan yang ditunjukkan dalam bulatan memang tidak sepadan dengan magnitud sebenar tentang sumber-sumber keberubahan berkenaan. Berdasarkan Rajah 3.7, setiap kesan utama dilambangkan dengan satu bentuk bulatan. Garis titik sesuatu bulatan

melambangkan faset pengukuran. Manakala kesan utama yang melibatkan sifat tersarang akan diwakili oleh satu bulatan yang berada dalam bulatan yang lain, misalnya kesan utama *r:t* diwakili oleh bulatan *r* yang berada dalam bulatan *t*.

Dalam kajian G, reka bentuk separa tersarang model rawak telah digunakan. Model ini tertakluk kepada dua andaian. Pertama, sampel calon, tugasan karangan dan pemeriksa adalah masing-masing dipilih secara rawak daripada populasi calon dan semesta tugasan dan pemeriksa, dan saiz untuk semua faset termasuk faset objek pengukuran adalah tak terhingga. Secara ringkas, $N \rightarrow \infty$ untuk semua faset termasuk faset objek pengukuran iaitu $N_p \rightarrow \infty$, $N_r \rightarrow \infty$, $N_t \rightarrow \infty$ (Brennan, 2001, p. 64) dan n < $N \rightarrow \infty$ (Brennan, 2001, p. 66). Kedua, korelasi di antara mana-mana dua kesan adalah kosong iaitu kesan skor mereka adalah tidak berkaitan sama sekali. Untuk populasi dan semesta, setiap komponen di sebelah kanan model tersebut merupakan satu variabel rawak kecuali min seluruh. Setiap variabel rawak pula mempunyai taburan tersendiri dan min taburannya adalah kosong (Brennan, 2001, p. 66). Ini dapat ditunjukkan seperti berikut:

$$E\,(v_p\, v_{p'}) = E\,(v_p\, v_t) = 0 \qquad (3.8)$$

di mana v_p dan $v_{p'}$ mewakili kondisi yang berbeza bagi kesan yang sama dan v_p dan v_t mewakili kesan yang berbeza (Brennan, 2001, p. 66). Selain itu, andaian reka bentuk teori G tidak menuntut andaian taburan normal bagi populasi dan semesta dan tidak juga mengandaikan bahawa kesan skor adalah bebas (Brennan, 2001, p. 24). Ciri kebebasan adalah andaian yang lebih kuat daripada andaian tidak ada korelasi.

3.6.1.2 Pengungkaian komponen varian

Berdasarkan ciri reka bentuk kajian dan andaian-andaian yang berkaitan, skor cerapan X_{prt} boleh dileraikan kepada lima komponen iaitu komponen varian calon $\sigma^2(p)$ dan tugasan $\sigma^2(t)$, interaksi dua hala komponen varian calon dengan tugasan $\sigma^2(pt)$, dan pemeriksa tersarang dalam tugasan $\sigma^2(r:t)$ serta interaksi tiga hala komponen varian calon dengan pemeriksa dengan tugasan $\sigma^2(pr:t)$. Dalam model ini, komponen varian kesan utama dan kesan bukan tersarang adalah seperti berikut:

$$\sigma^2(p) = \sigma^2(v_p) = Ev^2_{\mathrm{p}} = E(\mu_p - \mu)^2 = \sigma^2(\mu_p) \tag{3.9}$$

$$\sigma^2(t) = \sigma^2(v_t) = Ev^2_{\mathrm{t}} = E(\mu_t - \mu)^2 = \sigma^2(\mu_t) \tag{3.10}$$

$$\begin{aligned}\sigma^2(pt) = \sigma^2(v_{pt}) = Ev^2_{\mathrm{pt}} &= E(\mu_{pt} - \mu_p - \mu_t + \mu)^2 \\ &= E[(\mu_{pt} - \mu) - (\mu_p - \mu) - (\mu_t - \mu)]^2 \\ &= E[(\mu_{pt} - \mu) - v_p - v_t]^2\end{aligned} \tag{3.11}$$

Kesan interaksi $\sigma^2(pt)$ bagi *pt* dalam persamaan (3.11) boleh diertikan sebagai kesan setelah menyingkirkan kesan utama *p* dan *t* daripada $\boldsymbol{E}(\mu_{pt} - \mu)^2$ (Brennan, 2001, p. 75). Manakala komponen varian bagi kesan yang tersarang dalam faset lain adalah seperti berikut:

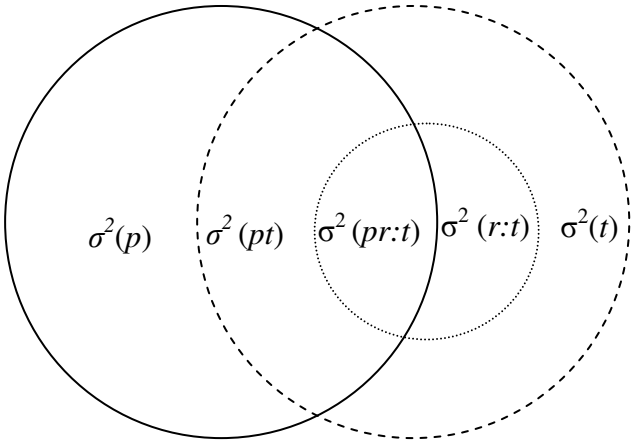
$$\sigma^2(r:t) = E(\mu_{rt} - \mu_t)^2 \tag{3.12}$$

$$\sigma^2(pr:t) = E(X_{prt} - \mu_{pt} - \mu_{rt} + \mu_t)^2 \tag{3.13}$$

Oleh itu, jumlah varian bagi skor cerapan X_{prt} boleh diungkaikan seperti berikut:

$$\sigma^2(X_{prt}) = \sigma^2(p) + \sigma^2(t) + \sigma^2(r:t) + \sigma^2(pt) + \sigma^2(pr:t) \tag{3.14}$$

Gambar rajah Venn adalah berguna untuk mengenal pasti sumber-sumber varian dalam pelbagai reka bentuk teori G. Pengungkaian komponen varian yang digambarkan dalam Rajah 3.8 adalah sejajar dengan pengungkaian kesan utama dalam Rajah 3.7.



Rajah 3.8. Gambar rajah Venn menunjukkan pengungkaian sumber-sumber komponen varian untuk reka bentuk separa tersarang dua faset $p \times (r{:}\ t)$. kesan rawak.

3.6.1.3 Anggaran komponen varian

Formula anggaran komponen varian untuk model $p \times (\ r{:}\ t)$ ditunjukkan dalam Jadual 3.15 dan Jadual 3.16. Perbincangan tentang formula-formula berkenaan adalah berasaskan pertimbangan populasi dan semesta. Dalam kerja pengukuran yang sebenar, penganggaran nilai-nilai tersebut memang tidak mungkin dilaksanakan. Namun begitu, kita boleh memperoleh nilai-nilai sampel bagi calon dan pemeriksa yang tersarang dalam tugasan karangan, dan menganggar parameter populasi dan semesta yang berkaitan berdasarkan nilai-nilai tersebut.

Pada masa kini, banyak program boleh digunakan untuk menganggar komponen varian iaitu seperti SAS, SPSS, BMDP, GENOVA (Brennan, 1994), LISREL, EQS dan AMOS (Marcoulides, 1996). Dalam kajian ini, program GENOVA telah digunakan untuk memproses data-data yang dikumpul untuk mendapatkan sumber-sumber komponen varian yang dikehendaki dan seterusnya mengoperasikan kajian D.

Jadual 3.15

Formula Bagi Anggaran Jumlah Kuasa Dua (SS) Dan Darjah Kebebasan (df) Untuk Setiap Kesan Dalam Reka Bentuk $p \times (r: t)$ Kajian G

Kesan (α)	df (α)	T(α)	SS (α)
p	$n_p - 1$	$n_r n_t \Sigma \overline{X}^2_p$	T(p) –T (μ)
t	$n_t - 1$	$n_p n_r \Sigma \overline{X}^2_t$	T(t) –T (μ)
$r:t$	$n_t (n_r - 1)$	$n_p \Sigma\Sigma \overline{X}^2_{r:t}$	T($r:t$) –T (t)
pt	$(n_p - 1) (n_t - 1)$	$n_r \Sigma\Sigma \overline{X}^2_{pt}$	T(pt) –T (p) –T(t) + T (μ)
$pr:t$	$n_t (n_p - 1) (n_r - 1)$	$\Sigma\Sigma\Sigma \overline{X}^2_{pr:t}$	T($pr:t$) –T (pt) –T($r:t$) + T (t)
Min (μ)		$n_p n_r n_t \overline{X}^2$	
Jumlah	$n_p n_r n_t - 1$		T($pr:t$) – T (μ)

Nota. α mewakili mana-mana kesan yang berkaitan.

Sumber: Brennan, R. L. (2001, p. 69-70, 432). *Generalizability theory*. New York: Springer-Verlag.

Jadual 3.16

Formula Anggaran Setiap Kesan Untuk Komponen-Komponen Varian Dalam Reka Bentuk p ×(r:t) Kajian G

Kesan (α)	EMS (α)	
p	$\sigma^2(pr{:}t) + n_r\sigma^2(pt) + n_r n_t\,\sigma^2(p)$	atau ss(*p*) / df(*p*)
t	$\sigma^2(pr{:}t) + n_r\sigma^2(pt) + n_p\,\sigma^2(r{:}t) + n_p n_r\sigma^2(t)$	atau ss(*t*) / df(*t*)
r:t	$\sigma^2(pr{:}t) + n_p\sigma^2(r{:}t)$	atau ss(*r:t*) / df(*r:t*)
pt	$\sigma^2(pr{:}t) + n_r\sigma^2(pt)$	atau ss(*pt*) / df(*pt*)
pr:t	$\sigma^2(pr{:}t)$	atau ss(*pr:t*)

Kesan (α)	σ^2 (α)
p	[MS(*p*) – MS(*pt*)] / $n_r n_t$
t	[MS(*t*) – MS(*r:t*) – MS(*pt*) + MS(*pr:t*)] / $n_p n_r$
r:t	[MS(*r:t*) – MS(*pr:t*)] / n_p
pt	[MS(*pt*) – MS(*pr:t*)] / n_r
pr:t	MS(*pr:t*)

Nota. α mewakili mana-mana kesan yang berkaitan. *EMS* (α) dan σ^2 (α) mewakili nilai jangkaan min kuasa dua dan nilai anggaran komponen varian masing-masing.

Sumber: Brennan, R. L. (2001, p. 76-82, 436). *Generalizability theory*. New York: Springer-Verlag.

3.6.2 Reka bentuk separa tersarang *p* × (*R:T*) model rawak dalam kajian D

Dalam kajian D, model reka bentuk kajian yang digunakan adalah sama dengan kajian G iaitu reka bentuk separa tersarang *p* ×(*R:T*) model rawak. Semua andaian dan definisi adalah sama bagi kedua-duanya kecuali kajian D meninjau min skor tentang kondisi set dalam semesta generalisasi dan bukan kondisi tunggal dalam semesta cerapan teraku seperti dalam kajian G (Brennan, 2001, p. 96). Sejajar dengan itu, indeks huruf kecil yang lazimnya digunakan bagi melambangkan setiap faset pengukuran dalam kajian G akan digantikan dengan huruf besar untuk mewakili setiap faset pengukuran semesta generalisasi dalam kajian D.

3.6.2.1 Model matematik bagi komponen varian

Secara matematik, pengungkaian min skor cerapan calon X_{pRT} dan kesan skor bagi reka bentuk separa tersarang $p \times (R:T)$ model kesan rawak boleh ditunjukkan seperti berikut:

Skor cerapan	Kesan skor
$X_{pRT} = \mu$	(μ = min seluruh bagi skor semesta calon)
$+ (\mu_p - \mu)$	(V_p= kesan calon)
$+ (\mu_T - \mu)$	(V_T = kesan bentuk tugasan)
$+ (\mu_{RT} - \mu_T)$	($V_{R:T}$ = kesan pemeriksa tersarang dalam tugasan)
$+ (\mu_{pT} - \mu_p - \mu_T + \mu)$	(V_{pT} = kesan interaksi calon dan tugasan)
$+ (X_{pRT} - \mu_{pT} - \mu_{RT} + \mu_T)$	($V_{pR:T}$ = kesan reja)

atau

$$X_{pRT} = \mu + \mu_p + \mu_T + \mu_{R:T} + \mu_{pT} + \mu_{pR:T} \quad (3.15)$$

$$= \mu + v_p + v_T + v_{R:T} + v_{pT} + v_{pR:T} \quad (3.16)$$

Min skor cerapan calon p, X_{pRT} merupakan nilai min bagi jumlah skor sampel yang diberi oleh mana-mana pemeriksa tersarang dalam tugasan karangan yang berlainan bentuk dalam kumpulan tertentu. Keadaan andaian yang tersirat bagi min skor cerapan X_{pRT} adalah seperti berikut:

$$\mu \equiv \underset{p}{E}\,\underset{R}{E}\,\underset{T}{E} X_{pRT} \tag{3.17}$$

$$\mu_T \equiv \underset{p}{E}\,\underset{R}{E} X_{pRT} \tag{3.18}$$

$$\mu_{RT} \equiv \underset{p}{E} X_{pRT} \tag{3.19}$$

$$= \underset{p}{E}(\mu + v_p + v_T + v_{R:T} + v_{pT} + v_{pR:T})$$

$$= \mu + \underset{p}{E}\, v_p + v_T + v_{R:T} + \underset{p}{E}\, v_{pT} + \underset{p}{E}\, v_{pR:T}$$

$$= \mu + v_T + v_{R:T} \text{ di mana } \boldsymbol{E}\,(v_\alpha) = 0 \text{ (Brennan, 2001: 65)}$$

$$\mu_p \equiv \underset{R}{E}\,\underset{T}{E} X_{pRT} \tag{3.20}$$

$$\mu_{pT} \equiv \underset{R}{E} X_{pRT} \tag{3.21}$$

3.6.2.2 Pengungkaian komponen varian

Seperti juga dalam kajian G, min skor cerapan X (*pRT*) dalam kajian D boleh dileraikan kepada lima komponen iaitu $\sigma^2(p)$, $\sigma^2(T)$, $\sigma^2(pT)$, $\sigma^2(R{:}T)$ dan $\sigma^2(pR{:}T)$. Dalam model ini, komponen varian kesan utama atau kesan bukan tersarang adalah seperti berikut:

$$\sigma^2(p) = \sigma^2(v_p) = Ev^2_p = E(\mu_p - \mu)^2 = \sigma^2(\mu_p) \tag{3.22}$$

$$\sigma^2(T) = \sigma^2(v_T) = Ev^2_T = E(\mu_T - \mu)^2 = \sigma^2(\mu_T) \tag{3.23}$$

$$\begin{aligned}\sigma^2(pT) = \sigma^2(v_{pT}) = Ev^2_{pT} &= E(\mu_{pT} - \mu_p - \mu_T + \mu)^2 \\ &= E\,[(\mu_{pT} - \mu) - (\mu_p - \mu) - (\mu_T - \mu)]^2 \\ &= E\,[(\mu_{pT} - \mu) - v_p - v_T]^2 \end{aligned} \tag{3.24}$$

Kesan interaksi $\sigma^2(pT)$ bagi *pT* dalam persamaan (3.24) boleh diertikan sebagai kesan setelah menyingkirkan kesan utama *p* dan *t* daripada $\boldsymbol{E}\,(\mu_{pT} - \mu)^2$. Manakala komponen varian bagi kesan yang tersarang dalam faset lain adalah seperti berikut:

$$\sigma^2(R{:}T) = E\,(\mu_{RT} - \mu_T)^2 \tag{3.25}$$

$$\sigma^2(pR{:}T) = E(X_{pRT} - \mu_{pT} - \mu_{RT} + \mu_T)^2 \tag{3.26}$$

Oleh itu, jumlah varian bagi min skor cerapan X_{pRT} (Brennan, 2001:15) boleh diungkaikan seperti berikut:

$$\sigma^2 (X_{pRT}) = \sigma^2(p) + \sigma^2 (T) + \sigma^2(R:T) + \sigma^2(pT) + \sigma^2(pR:T) \quad (3.27)$$

3.6.2.3 Anggaran komponen-komponen ralat

Ketidaktepatan generalisasi daripada skor cerapan berdasarkan sesuatu sampel pengukuran kepada skor semesta akan menghasilkan ralat pengukuran. Teori G membahagikan ralat pengukuran dalam kajian D kepada ralat relatif yang biasanya dilambangkan dengan simbol δ dan ralat mutlak yang diwakili oleh simbol Δ.

3.6.2.3.1 Ralat relatif dan ralat mutlak

Dalam Teori G, perbezaan telah dibuat antara ralat relatif dengan ralat mutlak. Ralat relatif diperlukan apabila interpretasi rujukan norma digunakan. Ia adalah sama dengan ralat dalam teori kebolehpercayaan klasik. Ralat relatif ditakrifkan sebagai perbezaan nilai min antara skor sisihan cerapan sampel sebenar calon dengan skor sisihan semesta sampel keseluruhan calon dalam semesta generalisasi (Brennan, 2001, p. 57). Simbol ralat relatif boleh ditulis sebagai δ_p atau δ_{pRT}. Bagi reka bentuk separa tersarang $p \times (R:T)$, ralat relatif (Brennan, 2001, p. 102) boleh digambarkan seperti berikut:

$$\delta_{pRT} = (\overline{X}_{pRT} - \mu_{RT}) - (\mu_p - \mu) \quad (3.28)$$

di mana μ_{RT} ialah nilai jangkaan $\overline{X}_{pRT}$ dan μ_p ialah skor semesta calon. Persamaan (3.28) boleh ditukar kepada:

$$\delta_{pRT} = (\mu_{pT} - \mu_p - \mu_T + \mu_{RT} + \mu) + (X_{pRT} - \mu_{pT} - \mu_{RT} + \mu_T + \mu_p + \mu_T + \mu_{RT})$$

iaitu,

$$\delta_{pRT} = v_{pT} + v_{pR:T} \quad (3.29)$$

Jika digambarkan dengan varian atau sisihan min kuasa dua ralat relatif, varian ralat relatif, $\sigma^2\delta_{pRT}$ adalah bersamaan dengan semua kesan interaksi komponen varian yang berkaitan dengan objek pengukuran dibahagikan dengan bilangan paras masing-masing (Brennan, 2001, p. 103), iaitu:

$$\sigma^2(\delta_{pRT}) = \sigma^2(pt) / \acute{n}_t + \sigma^2(pr{:}t) / \acute{n}_r\acute{n}_t \qquad (3.30)$$

$$= \sigma^2(pt) / \acute{n}_t + \sigma^2(pr) / \acute{n}_r\acute{n}_t + \sigma^2(prt) / \acute{n}_r\acute{n}_t$$

di mana $(pr{:}t) = \sigma^2(pr) + \sigma^2(prt)$ dan $\acute{n}_r\acute{n}_t$ merupakan bilangan sampel pemeriksa dan bentuk tugasan karangan dalam semesta generalisasi.

Ralat relatif merupakan kombinasi interaksi setiap faset dengan objek pengukuran. Justeru itu, kesan utama calon dan tugasan karangan serta kesan interaksi pemeriksa tersarang dalam tugasan karangan tidak akan mempengaruhi ralat relatif pengukuran. Ralat mutlak didefinisikan sebagai perbezaan nilai min antara skor cerapan dengan skor semesta calon pada semesta generalisasi (Brennan, 2001, p. 101). Ia digunakan apabila interpretasi rujukan kriteria diperlukan. Simbol ralat mutlak boleh ditulis dengan Δ_p atau Δ_{pRT}. Takrifan dan komponen varian ralat mutlak bagi reka bentuk $p \times (R{:}\ T)$ adalah seperti berikut:

$$\Delta_{pRT} \equiv \overline{X}_{pRT} - \mu_p \qquad (3.31)$$

iaitu

$$\Delta_{pRT} = v_T + v_{R:T} + v_{pT} + v_{pR:T} \qquad (3.32)$$

Menurut teori G, varian ralat mutlak terdiri daripada semua komponen varian kecuali kesan utama objek pengukuran dibahagikan dengan bilangan paras masing-masing (Brennan, 2001, p. 101). Varian ralat mutlak boleh ditakrifkan seperti berikut:

$$\sigma^2(\Delta_{pRT}) = \underset{p}{E}[\underset{R}{E}(\underset{T}{E}\,\Delta^2_{pRT})]$$

$$= \sigma^2(T) + \sigma^2(R{:}T) + \sigma^2(pT) + \sigma^2(pR{:}T) \qquad (3.33)$$

iaitu,

$$\sigma^2(\Delta) = \sigma^2(t)\ /\ \acute{n}_t + \sigma^2(r{:}t) / \acute{n}_r \acute{n}_t + \sigma^2(pt)\ /\ \acute{n}_t + \sigma^2(pr{:}t)\ /\ \acute{n}_r\acute{n}_t \qquad (3.34)$$

di mana $\acute{n}_r\acute{n}_t$ ialah bilangan sampel pemeriksa dan tugasan karangan pada semesta generalisasi. Formula di atas jelas menunjukkan bahawa dalam keputusan mutlak, semua komponen varian merupakan komponen varian ralat pengukuran kecuali kesan utama iaitu skor semesta calon.

Daripada perbandingan persamaan (3.30) dan (3.34), jelas bahawa nilai varian ralat relatif adalah lebih kecil atau sama dengan nilai varian ralat mutlak kerana semua komponen varian adalah positif.

3.6.2.3.2 Pekali G ($E\rho^2$) dan pekali phi (Φ)

Teori G membezakan antara pentafsiran keputusan tahap kedudukan relatif atau pangkatan calon dengan pentafsiran keputusan tahap kedudukan mutlak mengenai skor calon (Shavelson dan Webb, 1991, p. 84). Dalam hubungan ini, pekali G ($E\rho^2$) dalam teori G sesuai digunakan untuk membuat keputusan relatif berkaitan kebolehpercayaan skor calon dalam ujian rujukan norma. Ralat relatif dalam teori G boleh digambarkan dengan pekali G. Sekiranya tahap calon ialah objek pengukuran, maka definisi matematiknya boleh digambarkan seperti berikut:

$$E\rho^2 = \sigma^2(p) / [\sigma^2(p) + \sigma^2(pt) / \acute{n}_t + \sigma^2(pr{:}t) / \acute{n}_r\acute{n}_t\,]$$

$$= \sigma^2(p) / [\sigma^2(p) + \sigma^2(\delta)], \qquad (3.35)$$

di mana $\sigma^2(\delta) = \sigma^2(pt) / \acute{n}_t + \sigma^2(pr{:}t) / \acute{n}_r\acute{n}_t$ [lihat persamaan (3.30)]. Persamaan (3.35) membawa maksud bahawa nilai pekali G adalah terdiri daripada varian sah objek

pengukuran berbanding dengan jumlah varian sah objek pengukuran dan varian ralat relatif.

Manakala pekali phi atau indeks kebergantungan (Φ) dalam teori G sesuai digunakan untuk membuat keputusan mutlak berhubung kebolehpercayaan skor pemeriksa dalam ujian rujukan kriteria. Ralat mutlak dalam teori G boleh digambarkan dengan pekali phi. Sekiranya tahap calon ialah objek pengukuran, maka definisi matematiknya adalah seperti berikut:

$$\Phi = \sigma^2(p) \ / \ [\sigma^2(p) + \sigma^2(t) \ / \ \acute{n}_t + \sigma^2(pt) \ / \ \acute{n}_t + \sigma^2(r{:}t) \ / \ \acute{n}_r \acute{n}_t + \sigma^2(pr{:}t) \ / \ \acute{n}_r\acute{n}_t]$$

$$= \sigma^2(p) \ / \ [\sigma^2(p) + \sigma^2(\Delta)] \qquad (3.36)$$

Persamaan (3.36) membawa maksud bahawa nilai pekali phi adalah terdiri daripada varian sah objek pengukuran berbanding jumlah varian sah dan varian ralat mutlak.

Secara ringkas, pekali G hanya mempertimbangkan ralat rawak manakala pekali phi mengambil kira kedua-dua jenis ralat iaitu ralat rawak dan ralat sistematik. Justeru itu, untuk meninjau pengaruh ralat sistematik terhadap pemarkahan, kita perlu bergantung kepada pekali phi. Namun begitu, kajian ini hanya tertumpu kepada interpretasi rujukan norma iaitu membuat keputusan relatif yang melibatkan pekali G.

3.6.3 Model gabungan dalam kajian D

Selain reka bentuk separa tersarang $p \times (r{:}\ t)$ model rawak, model reka bentuk gabungan iaitu reka bentuk separa tersarang $p \times (r{:}\ t)$ dengan tugasan karangan sebagai faset tetap juga ditinjau dalam kajian D. Faset tetap merupakan semesta generalisasi yang dihadkan dan hanya melibatkan suatu subset tentang

kondisi dalam semesta cerapan teraku. Suatu faset dikatakan tetap dalam kajian D apabila $\acute{n} = \acute{N} < \infty$ dengan ń melambangkan saiz sampel suatu faset dalam kajian D dan Ń melambangkan saiz suatu faset dalam semesta generalisasi. Ini bermakna bilangan kondisi suatu faset dalam semesta generalisasi adalah terhingga iaitu $\acute{N} < \infty$, dan saiz sampel dalam kajian D menyamai bilangan kondisi suatu faset dalam semesta generalisasi iaitu $\acute{n} = \acute{N}$ (lihat Brennan, 2001, p. 121).

Berdasarkan reka bentuk model gabungan kajian D ini, penelitian dibuat dengan mengehadkan semesta generalisasi untuk faset tugasan. Ini bermakna tugasan yang sama akan digunakan bagi setiap kes prosedur pengukuran dalam semesta generalisasi. Selain itu, generalisasi akan dilakukan ke atas faset pemeriksa sahaja. Terdapat tiga komponen varian dalam model gabungan dengan *T* sebagai faset tetap (lihat Shavelson & Webb, 1991, p. 68; Brennan, 2001, p. 123), iaitu: $\sigma^2(p) = \sigma^2(p) + \sigma^2(pt) / \acute{n}_t$, $\sigma^2(R{:}T) = \sigma^2(r{:}t) / \acute{n}_r \acute{n}_t$ dan $\sigma^2(pR{:}T) = \sigma^2(pr{:}t) / \acute{n}_r\acute{n}_t$, di mana $\sigma^2(p)$ merupakan varian skor semesta, $\sigma^2(\tau)$, dalam model gabungan apabila $\acute{n}_t = n_t$.

Untuk faset *T* yang tetap, varian ralat relatif $[\sigma^2(\delta)]$ adalah seperti berikut (lihat Brennan, 2001, p. 123):

$$\sigma^2(\delta) = \sigma^2(pr{:}t) / \acute{n}_r\acute{n}_t$$
$$= \sigma^2(pr) / \acute{n}_r\acute{n}_t + \sigma^2(prt) / \acute{n}_r\acute{n}_t, \qquad (3.37)$$

di mana $(pr{:}t) = \sigma^2(pr) + \sigma^2(prt)$ dan $\acute{n}_r\acute{n}_t$ merupakan bilangan sampel pemeriksa dan bentuk tugasan karangan dalam semesta generalisasi. Apabila faset *T* ditetapkan, komponen varian $\sigma^2(pT)$ akan menyumbang kepada varian skor semesta dan bukan kepada varian ralat relatif $\sigma^2(\delta)$ lagi. Justeru itu, nilai $\sigma^2(\delta)$ bagi reka bentuk separa tersarang $p \times (r{:}\ t)$ dengan faset *T* tetap adalah lebih kecil berbanding dengan model

rawak reka bentuk yang sama [lihat persamaan (3.30)]. Ini juga bermakna varian skor semestanya adalah lebih besar daripada varian skor semesta dalam model rawak. Manakala varian ralat mutlak [$\sigma^2(\Delta)$] adalah seperti berikut:

$$\sigma^2(\Delta) = \sigma^2(r{:}t) \,/\, \acute{n}_r \acute{n}_t + \sigma^2(pr{:}t) \,/\, \acute{n}_r \acute{n}_t, \tag{3.38}$$

iaitu dengan faset *T* ditetapkan, $\sigma^2(\Delta)$ tidak lagi mengandungi $\sigma^2(T)$ kerana setiap kes dalam prosedur pengukuran telah mengandungi kondisi $\acute{n}_t$ yang sama manakala $\sigma^2(pT)$ merupakan sebahagian daripada varian skor semesta [bandingkan dengan persamaan (3.34)]. Berdasarkan nilai varian ralat relatif dan varian mutlak, anggaran pekali G ($E\rho^2$) dan pekali phi (Φ) boleh ditunjukkan seperti berikut:

$$E\rho^2 = \sigma^2(\tau) \,/\, [\sigma^2(\tau) + \sigma^2(\delta)] \text{ [lihat persamaan (3.35)]}$$

$$= \frac{\sigma^2(p) + \sigma^2(pt) \,/\, \acute{n}_t}{\sigma^2(p) + \sigma^2(pt) \,/\, \acute{n}_t + \sigma^2(pr) \,/\, \acute{n}_r \acute{n}_t + \sigma^2(prt) \,/\, \acute{n}_r} \tag{3.39}$$

$$\Phi = \sigma^2(p) \,/\, [\sigma^2(p) + \sigma^2(\Delta)] \text{ [lihat persamaan (3.36)]}$$

$$= \frac{\sigma^2(p) + \sigma^2(pt) \,/\, \acute{n}_t}{\sigma^2(p) + \sigma^2(pt) \,/\, \acute{n}_t + \sigma^2(r{:}t) \,/\, \acute{n}_r \acute{n}_t + \sigma^2(pr{:}t) \,/\, \acute{n}_r \acute{n}_t} \tag{3.40}$$

di mana $\sigma^2(\delta) = \sigma^2(pr) \,/\, \acute{n}_r \acute{n}_t + \sigma^2(prt) \,/\, \acute{n}_r \acute{n}_t$, [lihat persamaan (3.37)] dan $\sigma^2(\Delta) = \sigma^2(r{:}t) \,/\, \acute{n}_r \acute{n}_t + \sigma^2(pr{:}t) \,/\, \acute{n}_r \acute{n}_t$ [lihat persamaan (3.38)]. Bagaimanapun, kajian ini melibatkan membuat keputusan relatif, oleh itu hanya pekali G dilaporkan.

3.7 Tata Cara Kajian

Sebelum menjalankan penyelidikan ini, pengkaji telah menemu dan berbincang dengan penyelia tesis dan pakar-pakar pentaksiran tentang bidang kajian yang hendak dilakukan. Seterusnya pengkaji membuat rujukan daripada bahan-bahan sekunder sama ada kajian tempatan atau luar negara di perpustakaan, pusat sumber dan melalui internet. Bahan-bahan tersebut ditelaah dan dikaji untuk menyediakan

cadangan kajian. Pengkaji juga mendapat tunjuk ajar dan nasihat daripada penyelia dari semasa ke semasa dalam proses untuk menyiapkan cadangan kajian.

Setelah mengenal pasti lokasi dan sampel kajian, pengkaji berusaha untuk mendapatkan senarai populasi murid, bilangan kelas dan keputusan ujian penulisan UPSR untuk tiga tahun berturut-turut daripada pihak pentadbir sekolah daerah kajian pada awal tahun 2008. Di samping itu, pengkaji telah memohon soalan-soalan ujian karangan Bahasa Cina tahun-tahun lepas yang digunakan dalam penilaian daerah dan negeri daripada Jabatan Pelajaran Perak melalui Jabatan Psikologi Pendidikan dan Kaunseling, Fakulti Pendidikan, Universiti Malaya. Pengkaji telah mendapat bantuan empat orang guru pakar mata pelajaran Bahasa Cina sekolah rendah untuk menilai dan seterusnya memilih soalan-soalan karangan bagi tujuan kajian rintis.

Selain itu, pengkaji juga meminta bantuan tiga orang Guru Cemerlang yang juga merupakan jurulatih utama PPK dalam mata pelajaran Bahasa Cina untuk memurni dan memantapkan draf-draf instrumen pemarkahan. Untuk tujuan penilaian pakar pentaksiran, instrumen-instrumen berkenaan diterjemahkan ke dalam Bahasa Melayu oleh tiga orang guru Bahasa Melayu dan kemudian diterjemahkan semula ke dalam Bahasa Cina oleh tiga orang guru Bahasa Cina yang lain dengan menggunakan kaedah *back translation*. Seterusnya, pengkaji meminta tiga orang pakar pentaksiran daripada badan peperiksaan dalam negara untuk menilai keberkesanan instrumen-instrumen berkenaan.

Untuk menentukan kebolehlaksanaan soalan karangan dan ketekalan instrumen pemarkahan, pengkaji telah merancang untuk menjalankan kajian rintis.

Pengkaji telah berbincang dengan penyelia tentang kajian rintis dan soal selidik yang hendak dijalankan. Setelah mengenal pasti sasaran sekolah dari segi kesesuaian ciri populasi dengan kajian sebenar, pengkaji membuat temu janji dengan salah sebuah sekolah SJK(C) yang dipilih untuk menetapkan tarikh bagi mengendalikan kajian rintis. Pengkaji juga mendapat kebenaran daripada pihak sekolah untuk menggunakan bilik darjah dan kemudahan lain semasa menjalankan penyelidikan. Selain menguji soalan karangan, pengkaji juga menjalankan kajian soal selidik terhadap soalan karangan tersebut ke atas murid Tahun Enam. Skrip jawapan karangan murid telah dinilai oleh pemeriksa berpengalaman dengan berdasarkan instrumen pemarkahan yang disediakan setelah latihan pemarkahan diberikan. Skor karangan dan data kajian soal selidik dalam kajian rintis telah dianalisis. Dengan bimbingan dan tunjuk ajar penyelia, soalan karangan yang digunakan dimurnikan seterusnya agar sesuai dan menepati objektif kajian.

Selepas itu, pengkaji mengemukakan satu kertas cadangan penyelidikan kepada Bahagian Perancangan dan Penyelidikan Pendidikan (BPPP), Kementerian Pelajaran Malaysia (KPM) untuk memohon surat kebenaran bagi menjalankan penyelidikan di sekolah-sekolah di bawah KPM (Lampiran O). Setelah mendapat kelulusan daripada BPPP, pengkaji memohon surat kebenaran daripada pengarah JPN Perak untuk menggunakan sampel kajian di SJK(C) di sebuah daerah di Perak (Lampiran P). Pada masa yang sama, pengkaji berusaha mendapat 12 orang pemeriksa untuk memeriksa skrip jawapan murid.

Selepas itu, pengkaji meminta kebenaran daripada pihak pentadbir sekolah sebelum melawat ke sekolah-sekolah tersebut. Semasa berurusan dengan pihak

sekolah, pengkaji menampilkan surat kebenaran daripada BPPP dan JPP kepada pihak pentadbir. Seterusnya, pengkaji menerangkan rasional dan tujuan penyelidikan yang hendak dijalankan serta berbincang dengan pihak pentadbir cara-cara mengendalikan penyelidikan tersebut.

Pengkaji juga meminta pihak pentadbir mencadangkan seorang guru yang sudi menjadi pembantu kajian, kecuali SJK(C) B2 dan B4 memerlukan dua orang guru, untuk mengendalikan kajian ini. Kriteria-kriteria pembantu kajian yang diperlukan adalah pernah dilantik sebagai pengawas UPSR dan berpengalaman mengajar sekurang-kurangnya selama 10 tahun agar mampu melaksanakan tugas pengawasan tersebut. Seminggu sebelum pentadbiran kajian dijalankan, pengkaji telah mengadakan satu mesyuarat dengan pembantu-pembantu kajian untuk membincang dan menerangkan cara-cara mentadbirkan ujian penulisan di salah sebuah sekolah di daerah tersebut. Mereka juga diminta menjelaskan rasional dan tujuan penyelidikan kepada sampel kajian yang terlibat iaitu murid Tahun Enam. Kajian tersebut ditadbirkan secara serentak di sepuluh buah kelas yang terdapat di lapan buah sekolah dalam daerah tersebut. Pentadbiran kajian dijalankan sebanyak dua kali dalam jangka masa dua minggu iaitu seminggu sekali.

Pengkaji juga mengagihkan soalan-soalan karangan, kertas tulis, arahan ujian karangan dan bahan-bahan yang diperlukan semasa pengendalian ujian kepada pembantu kajian di sekolah-sekolah yang terlibat dalam mesyuarat tersebut. Soalan-soalan tersebut disimpan dalam sampul surat dan digamkan. Pembantu kajian diminta menyimpan soalan kajian di tempat yang selamat dan berkunci untuk mengelakkan kebocoran. Pengkaji mengutip skrip jawapan daripada pembantu kajian

selepas sahaja pentadbiran ujian selesai dijalankan. Jumlah skrip karangan yang dikumpul daripada 10 buah kelas yang melibatkan 8 buah sekolah dalam kajian ini ialah 704 iaitu 353 skrip pada ujian kali pertama dan 351 skrip pada ujian kali kedua (Jadual 3.17).

Berkenaan dengan pemeriksaan skrip jawapan calon, pengkaji telah mendapatkan kebenaran pihak sekolah untuk menggunakan bilik mesyuarat sekolah sebagai pusat pemarkahan. Selain itu, skrip jawapan telah diperbanyakkan dan dimasukkan dalam sampul yang berasingan untuk kegunaan pemeriksaan kumpulan pemeriksa yang berlainan. Prosedur-prosedur latihan pemeriksa dan kerja pemeriksaan skrip karangan telah dijelaskan dalam fasal 3.4.8 iaitu pemarkahan sampel karangan. Data yang dikutip dianalisis dengan program SPSS dan Genova.

Jadual 3.17

Bilangan Skrip Karangan Yang Dikumpul

	Bil Kelas		Bil Skrip Dikumpul	
Sekolah	Terpilih	Bil Murid	Ujian 1	Ujian 2
SJK(C) B1	1	38	37	36
SJK(C) B2	2	72	70	71
SJK(C) B3	1	39	38	38
SJK(C) B4	2	82	80	79
SJK(C) B5	1	30	30	29
SJK(C) L1	1	32	31	30
SJK(C) L2	1	37	35	36
SJK(C) L3	1	33	32	32
Jumlah	10	363	353	351

3.8 Analisis Data

Dalam kajian ini, markah bagi dua tugasan karangan dikira berdasarkan empat prosedur pemarkahan yang berlainan iaitu (1) kaedah holistik serta aspek kandungan dan organisasi; (2) kaedah analitik serta aspek kandungan dan organisasi; (3) kaedah holistik serta aspek penggunaan bahasa dan mekanis; dan (4) kaedah analitik serta aspek penggunaan bahasa dan mekanis. Prosedur pemarkahan tersebut dinilai oleh kumpulan pemeriksa yang berlainan yang setiap satu terdiri daripada tiga orang pemeriksa. Menurut jangkaan teori ujian klasik dan teori G, pengendalian skor berdasarkan purata skor yang diberikan oleh tiga orang pemeriksa mempunyai kebolehpercayaan yang lebih tinggi (Johnson, Penny, & Gordon, 2000).

Untuk kaedah holistik, setiap aspek pemarkahan mempunyai lima band dengan setiap satu band diperuntukan sebanyak dua markah. Oleh itu, jumlah skor bagi setiap prosedur pemarkahan ialah 10 markah. Keadaan agak berlainan sedikit bagi pemarkahan dengan menggunakan kaedah analitik. Markah untuk aspek pemarkahan iaitu aspek kandungan dan organisasi telah dibahagikan kepada dua bahagian dan dinilai secara berasingan. Setiap bahagian juga mempunyai lima band tetapi setiap band hanya diperuntukan satu markah. Ini bermakna jumlah skornya bagi setiap bahagian ialah lima markah. Setelah penilaian, markah bagi kedua-dua bahagian tersebut dijumlahkan dan dijadikan satu skor tunggal. Aspek penggunaan bahasa dan mekanis juga menggunakan cara perhitungan yang sama. Jadual 3.18 dan Jadual 3.19 secara ringkas menunjukkan cara-cara pengiraan markah bagi prosedur pemarkahan yang berlainan berdasarkan karangan berbeza bentuk agar mudah difahami.

Jadual 3.18

Pengagihan Markah Untuk Karangan Berlainan Tugasan Berdasarkan Kaedah Holistik Dengan Aspek Pemarkahan Yang Berlainan

Prosedur Pemarkahan	Tugasan Karangan	Band	Jumlah Markah	Cara Penilaian
Kaedah Holistik Serta Kandungan Dan Organisasi	Naratif	5	10 (5x2)	Kumpulan pemeriksa yang berlainan (3 orang bagi setiap kumpulan)
	Pendedahan			
Kaedah Holistik Serta Penggunaan Bahasa Dan Mekanis	Naratif			
	Pendedahan			

Jadual 3.19

Pengagihan Markah Untuk Karangan Berlainan Tugasan Berdasarkan Kaedah Analitik Dengan Aspek Pemarkahan Yang Berlainan

Prosedur Pemarkahan		Tugasan Karangan	Band	Jumlah Markah	Cara Penilaian
Kaedah Analitik	Kandungan	Naratif	5	10	Kumpulan pemeriksa yang berlainan (3 orang bagi setiap kumpulan)
	Organisasi		5		
	Kandungan	Pendedahan	5	10	
	Organisasi		5		
	Penggunaan Bahasa	Naratif	5	10	
	Mekanis		5		
	Penggunaan Bahasa	Pendedahan	5	10	
	Mekanis		5		

Markah atau data yang diperoleh daripada kedua-dua buah karangan berbeza bentuk berdasarkan prosedur pemarkahan yang berlainan adalah penting untuk mendapat maklumat tentang min dan sisihan piawai bagi skor karangan, kebolehpercayaan antara pemeriksa, ujian *Levene* dan ujian *t*, dan analisis *generalizability* dalam kajian G dan kajian D. Data-data yang dikumpulkan dalam

kajian ini diproses dengan program *Statistical Packages for the Social Sciences* (*SPSS*) kecuali analisis *generalizability*. Maklumat pengendalian program *SPSS* yang disarankan oleh Norusis (1999) dan Coakes (2005) dijadikan sebagai panduan.

Dalam kajian ini, statistik deskriptif iaitu min dan sisihan piawai digunakan untuk menganalisis lapan jenis keadaan pemarkahan yang berlainan (dua aspek pemarkahan × dua kaedah pemarkahan × dua tugasan karangan) terhadap 120 calon. Setiap jenis keadaan dinilai oleh tiga orang pemeriksa yang berlainan. Statistik inferensi pula digunakan untuk menganalisis data bagi menghasilkan pekali korelasi antara pemeriksa (tiga orang dalam setiap kumpulan) berdasarkan karangan berbeza bentuk mengikut prosedur pemarkahan yang berlainan. Pekali korelasi yang tinggi bermakna kebolehpercayaan skor yang diberi oleh pemeriksa yang berlainan dalam setiap kumpulan adalah tinggi.

Statistik inferensi iaitu ujian *t* untuk kumpulan-kumpulan bebas (*independent sampel t-test)* pula digunakan untuk melihat kesignifikanan karangan berbeza bentuk berdasarkan prosedur pemarkahan yang berlainan. Keputusan ujian *Levene* yang didapati adalah untuk mengesan kesetaraan varian bagi kedua-dua tugasan karangan dalam kajian ini. Dapatan ujian tersebut juga adalah untuk menentukan nilai *t* yang sesuai digunakan. Sekiranya keputusan ujian *Levene* adalah signifikan ($p < .05$), maka baris kedua dalam jadual ujian *t* iaitu varian setara tidak diandaikan dirujuk. Sebaliknya, sekiranya ujian tersebut adalah tidak signifikan ($p > .05$), maka ini bermakna kedua-dua kumpulan adalah homogen dan maklumat baris pertama iaitu varian setara diandaikan dirujuk. Statistik ujian *t* pada paras keyakinan .05 digunakan dalam menguji hipotesis kajian. Ini bermakna peluang untuk melakukan kesilapan

adalah lima peratus dan piawaian tersebut biasanya diamalkan dalam penyelidikan sains sosial (Bahaman Abu Samad & Turiman Suandi, 1999). Untuk menguji hipotesis kajian yang dikemukakan, ujian *t* dua hujung digunakan untuk mencari sama ada terdapat perbezaaan antara karangan bentuk naratif dengan pendedahan berdasarkan empat prosedur pemarkahan yang berbeza. Ujian dua hujung digunakan kerana hipotesis yang dinyatakan tidak menunjukkan arah perbezaan sama ada ke arah positif atau ke arah negatif. Menurut Sidek Mohd. Noah (2002), ujian dua hujung sesuai digunakan apabila penyelidik ingin menentukan sama ada rawatan mendatangkan kesan, sama ada meningkatkan atau menurunkan secara signifikan min sampel. Pemilihan statistik untuk menguji empat hipotesis yang dibentuk bersama variabel bersandar dan variabel tak bersandar yang digunakan dalam kajian ini telah ditunjukkan dalam Jadual 3.20.

Untuk menentukan kesan saiz (*effect size*) bagi ujian *t*, *eta squared* (η^2) telah digunakan. Kesan saiz adalah penunjuk tentang perbezaan dari segi magnitud antara kumpulan. Nilai *eta squared* adalah di antara 0 hingga 1. Nilai tersebut digunakan untuk menghuraikan perubahan amaun varian dalam variabel bersandar yang disebabkan oleh variabel tak bersandar. Ini bermakna sekiranya kesan saiz lebih besar, maka amaun variabel bersandar yang meliputi variabel tak bersandar akan bertambah atau keadaan sebaliknya akan berlaku. Menurut Cohen (1988), nilai η^2 yang kurang daripada .06 menunjukkan kesan yang kecil, nilai yang lebih besar daripada .06 tetapi kurang daripada .14 menunjukkan kesan yang sederhana kuat manakala nilai yang lebih besar daripada .14 menandakan kesan yang kuat.

Jadual 3.20

Penganalisisan Data Statistik

Bil	*Hipotesis*	*DV*	*IV*	*Statistik*
1.	Terdapat perbezaaan yang signifikan antara karangan bentuk naratif dengan karangan bentuk pendedahan berdasarkan prosedur pemarkahan kaedah holistik serta aspek kandungan dan organisasi.	Skor karangan (Sela)	Karangan bentuk naratif dan pendedahan (Nominal)	Ujian *t* (jenis dua hujung)
2.	Terdapat perbezaaan yang signifikan antara karangan bentuk naratif dengan karangan bentuk pendedahan berdasarkan pemarkahan kaedah kaedah analitik serta aspek kandungan dan organisasi.	Skor karangan (Sela)	Karangan bentuk naratif dan pendedahan (Nominal)	Ujian *t* (jenis dua hujung)
3.	Terdapat perbezaaan yang signifikan antara karangan bentuk naratif dengan karangan bentuk pendedahan berdasarkan pemarkahan kaedah kaedah holistik serta aspek penggunaan bahasa dan makanis.	Skor karangan (Sela)	Karangan bentuk naratif dan pendedahan (Nominal)	Ujian *t* (jenis dua hujung)
4.	Terdapat perbezaaan yang signifikan antara karangan bentuk naratif dengan karangan bentuk pendedahan berdasarkan pemarkahan kaedah kaedah analitik serta aspek penggunaan bahasa dan makanis.	Skor karangan (Sela)	Karangan bentuk naratif dan pendedahan (Nominal)	Ujian *t* (jenis dua hujung)

Nota. *DV* = variabel bersandar; *IV* = variabel tak bersandar.

Selain program *SPSS*, program GENOVA (Crick dan Brennan, 1983) yang direka khas bagi memenuhi penggunaan teori G dalam analisis varian juga digunakan. Program GENOVA bersesuaian dengan reka bentuk ANOVA yang lengkap (*complete*) dan seimbang. Program tersebut akan digunakan untuk mengungkai dan menganggar komponen varian dalam kajian G, dan juga menghasilkan pekali G berdasarkan kondisi pengukuran yang ditetapkan dalam

kajian ini. Selain itu, program tersebut juga digunakan untuk meninjau perubahan pekali G dalam kajian susulan iaitu kajian D berdasarkan bilangan karangan dan bilangan pemeriksa serta kombinasi kedua-dua faset berkenaan dalam prosedur pengukuran.

Dalam kajian ini, nilai pekali kebolehpercayaan .90 digunakan untuk menandakan tahap kebolehpercayaan yang tinggi. Keperluan darjah sesuatu nilai pekali kebolehpercayaan dalam pentaksiran pendidikan banyak bergantung kepada keputusan yang hendak dibuat. Menurut Linn dan Gronlund (2005), ujian buatan guru biasanya mempunyai pekali kebolehpercayaan antara .60 dan .85. Menurut Nunnally dan Burnstein (1994), nilai pekali kebolehpercayaan .80 adalah tidak mencukupi untuk tujuan membuat keputusan tentang individu dan untuk membuat keputusan yang penting berkaitan dengan skor ujian yang spesifik, nilai pekali kebolehpercayaan .90 dianggap sebagai keperluan minimum (Swartz, et al., 1999). Secara terperinci, nilai pekali kebolehpercayaan .90 menunjukkan bahawa kira-kira 32% (punca kuasa dua .10 = .316) daripada sisihan piawai bagi skor cerapan adalah dipengaruhi oleh ralat pengukuran (Pawlik, 2003, p.1021). Berdasarkan pandangan pakar-pakar pentaksiran tersebut, pengkaji telah menetapkan kriteria .90 sebagai penandaarasan (*bench mark*) untuk menunjukkan pekali kebolehpercayaan yang tinggi bagi membuat keputusan relatif berdasarkan data kajian ini.

3.9 Rumusan Bab

Bab ini telah membentangkan kaedah yang digunakan dalam menjalankan kajian ini. Aspek-aspek yang telah dibincangkan meliputi reka bentuk kajian, lokasi kajian, sampel kajian, instrumen kajian, tata cara kajian dan analisis data.

BAB IV

KEPUTUSAN KAJIAN

4.0 Pengenalan

Bab ini membentangkan hasil kajian tentang keputusan statistik deskriptif yang melibatkan min dan sisihan piawai bagi setiap tugasan karangan dan keputusan statistik inferensi untuk menguji perbezaan skor dan kesetaraan varian bagi tugasan karangan berlainan. Analisis keputusan komponen-komponen varian bagi kajian G berdasarkan prosedur pemarkahan yang berlainan telah dikemukakan. Selain itu, analisis keputusan kajian D yang berkaitan dengan kebergantungan skor karangan di bawah prosedur pemarkahan yang berlainan juga dibentangkan.

4.1 Analisis Keputusan Dalam Kajian G

4.1.1 Keputusan kebolehpercayaan antara pemeriksa dan statistik deskriptif

Jadual 4.1 memaparkan pekali korelasi bagi tiga orang pemeriksa dalam pemarkahan tugasan karangan tertentu berdasarkan prosedur pemarkahan yang berlainan. Pemarkahan tersebut dijalankan sebanyak dua kali secara berasingan. Nilai yang tinggi bagi kebolehpercayaan antara pemeriksa dianggap sebagai bukti bahawa jumlah pemarkahan yang dilakukan ke atas calon dapat mencerminkan dengan tepat prestasi calon (Linacre, 1999, p.246). Pada keseluruhannya, keputusan pekali kebolehpercayaan yang ditunjukkan adalah agak memuaskan iaitu semuanya berada di atas .80 dengan aras signifikan $p < .01$. Ini bermakna bahawa sekurang-kurangnya

80% daripada varian skor adalah boleh dipercayai dan yang selebihnya merupakan ralat pengukuran (Brown, 2005, p.175). Anggaran pekali kebolehpercayaan tersebut menunjukkan skor yang diberikan oleh pemeriksa yang berlainan berdasarkan tugasan karangan tertentu di bawah prosedur pemarkahan yang berlainan adalah boleh dipercayai.

Jadual 4.1

Kebolehpercayaan Antara Pemeriksa Bagi Tugasan Karangan Berlainan Berdasarkan Prosedur Pemarkahan Yang Berbeza (N = 120)

Prosedur Pemarkahan	Pemarkahan Kali Pertama			Prosedur Pemarkahan	Pemarkahan Kali Kedua		
nhk	R1	R2	R3	pab	R1	R2	R3
R1	1.000			R1	1.000		
R2	.814**	1.000		R2	.850**	1.000	
R3	.810**	.806**	1.000	R3	.814**	.802**	1.000
phk	R4	R5	R6	nab	R4	R5	R6
R4	1.000			R4	1.000		
R5	.831**	1.000		R5	.836**	1.000	
R6	.813**	.822**	1.000	R6	.801**	.805**	1.000
nak	R7	R8	R9	phb	R7	R8	R9
R7	1.000			R7	1.000		
R8	.819**	1.000		R8	.841**	1.000	
R9	.805**	.827**	1.000	R9	.873**	.885**	1.000
pak	R10	R11	R12	nhb	R10	R11	R12
R10	1.000			R10	1.000		
R11	.804**	1.000		R11	.802**	1.000	
R12	.815**	.849**	1.000	R12	.831**	.810**	1.000

Nota. nhk = karangan berunsur naratif berdasarkan pemarkahan kaedah holistik serta aspek kandungan dan organisasi; phk = karangan berunsur pendedahan berdasarkan pemarkahan kaedah holistik serta aspek kandungan dan organisasi; nak = karangan berunsur naratif berdasarkan pemarkahan kaedah analitik serta aspek kandungan dan organisasi; pak = karangan berunsur pendedahan berdasarkan pemarkahan kaedah analitik serta aspek kandungan dan organisasi; pab = karangan berunsur pendedahan berdasarkan pemarkahan kaedah analitik serta aspek penggunaan bahasa dan mekanis; nab = karangan berunsur naratif berdasarkan pemarkahan kaedah analitik serta aspek penggunaan bahasa dan mekanis; phb = karangan berunsur pendedahan berdasarkan pemarkahan kaedah holistik serta aspek penggunaan bahasa dan mekanis; nhb = karangan berunsur naratif berdasarkan pemarkahan kaedah holistik serta aspek penggunaan bahasa dan mekanis.
R1 = pemeriksa pertama; R2 = pemeriksa kedua; R3 = pemeriksa ketiga dan seterusnya.

$^{**}p < .01$.

Jadual 4.2 menunjukkan keputusan min dan sisihan piawai bagi 120 subjek yang dinilai berdasarkan lapan jenis keadaan kajian (dua tugasan × dua kaedah pemarkahan × dua aspek pemarkahan). Keputusan ini diperoleh melalui pemarkahan oleh empat kumpulan pemeriksa yang tersarang dalam tugasan karangan berlainan dengan setiap kumpulan mengandungi tiga orang pemeriksa. Setiap kumpulan pemeriksa telah menilai kedua-dua buah karangan berdasarkan aspek dan kaedah pemarkahan yang berlainan.

Jadual 4.2

Keputusan Min Dan Sisihan Piawai Bagi Tugasan Karangan Berlainan Berdasarkan Aspek Dan Kaedah Pemarkahan Yang Berlainan

		Aspek Kandungan Dan Organisasi		Aspek Penggunaan Bahasa Dan Mekanis	
Tugasan Karangan		Kaedah Holistik	Kaedah Analitik	Kaedah Holistik	Kaedah Analitik
Naratif	*M*	6.60	6.62	6.88	7.05
	SD	1.424	1.495	1.493	1.375
Pendedahan	*M*	5.85	6.46	6.40	6.91
	SD	1.871	1.726	1.799	1.719

Jika dilihat dari segi aspek pemarkahan, sisihan piawai bagi kedua-dua buah tugasan karangan yang ditaksir dengan aspek kandungan dan organisasi masing-masing adalah 1.424, 1.495, 1.726 dan 1.871. Manakala bagi pentaksiran aspek penggunaan bahasa dan mekanis pula, sisihan piawai bagi tugasan karangan berlainan tersebut adalah masing-masing 1.375, 1.493, 1.719 dan 1.799. Ini menunjukkan bahawa sisihan piawai bagi pentaksiran aspek kandungan dan organisasi adalah lebih tinggi daripada pentaksiran aspek penggunaan bahasa dan

mekanis. Dapatan keputusan kajian ini telah membuktikan bahawa keberubahan skor bagi kaedah pemarkahan yang menggunakan aspek kandungan dan organisasi adalah lebih tinggi daripada aspek penggunaan bahasa dan mekanis.

Jika dilihat dari segi kaedah pemarkahan, sisihan piawai bagi kedua-dua buah tugasan karangan yang ditaksir dengan kaedah holistik masing-masing adalah 1.424, 1.493, 1.799 dan 1.871. Manakala bagi kaedah analitik pula, sisihan piawai bagi tugasan karangan berlainan tersebut adalah masing-masing 1.375, 1.495, 1.719 dan 1.726. Ini menunjukkan bahawa sisihan piawai bagi kaedah holistik adalah lebih tinggi daripada kaedah analitik. Keputusan ini mungkin menunjukkan keberubahan skor bagi pemarkahan yang menggunakan kaedah holistik adalah lebih tinggi daripada kaedah analitik.

Selain itu, karangan berunsur naratif yang ditaksir dengan kaedah analitik serta aspek penggunaan bahasa dan mekanis memperoleh min yang paling tinggi manakala sisihan piawainya adalah paling rendah iaitu masing-masing 7.05 dan 1.375. Tetapi keadaan adalah sebaliknya bagi pentaksiran karangan berunsur pendedahan yang menggunakan pemarkahan kaedah holistik serta aspek kandungan dan organisasi, minnya adalah paling rendah manakala sisihan piawainya adalah paling tinggi iaitu masing-masing 5.85 dan 1.871.

Analisis data kajian di atas telah menggambarkan bahawa tugasan karangan berlainan, pemarkahan pemeriksa, kaedah pemarkahan dan aspek pemarkahan boleh menyebabkan keberubahan skor. Kesannya adalah agak jelas dalam statistik inferensi yang dikupaskan selanjutnya.

4.1.2 Keputusan berkaitan dengan statistik inferensi

Untuk melihat keadaan perbezaan skor antara karangan berunsur naratif dengan karangan berunsur pendedahan berdasarkan empat jenis prosedur pemarkahan yang berlainan iaitu (1) kaedah holistik serta aspek kandungan dan organisasi, (2) kaedah analitik serta aspek kandungan dan organisasi, (3) kaedah holistik serta aspek penggunaan bahasa dan mekanis, dan (4) kaedah analitik serta aspek penggunaan bahasa dan mekanis, ujian *t* telah digunakan.

Jadual 4.3 hingga Jadual 4.6 adalah hasil kajian yang menunjukkan keputusan analisis kesignifikanan min bagi tugasan karangan berlainan berdasarkan prosedur pemarkahan yang berlainan dengan menggunakan ujian *t*. Dapatan keputusan jelas menunjukkan bahawa kedua-dua buah tugasan karangan yang dinilai dengan prosedur pemarkahan yang berlainan menghasilkan keputusan yang berbeza. Perbezaan min skor bagi karangan berunsur naratif dan karangan berunsur pendedahan adalah signifikan berdasarkan pemarkahan kaedah holistik serta aspek kandungan dan organisasi [$t(670.480) = 6.097$, $p = .000 < .05$], dan kaedah holistik serta aspek penggunaan bahasa dan mekanis [$t(694.414) = 3.877$, $p = .000 < .05$] (Jadual 4.3 dan Jadual 4.5). Walau bagaimanapun, nilai *eta* kuasa dua (η^2) antara tugasan karangan berlainan dengan skor karangan berdasarkan kedua-dua prosedur pemarkahan tersebut masing-masing adalah .049 dan .021. Ini menunjukkan bahawa tugasan karangan berlainan hanya meliputi 4.9% dan 2.1% daripada varian skor karangan dan kesan saiz tugasan karangan ke atas skor karangan adalah kecil (Cohen, 1988). Manakala perbezaan min skor untuk tugasan karangan berlainan bagi dua lagi prosedur pemarkahan yang lain iaitu pemarkahan kaedah analitik serta aspek kandungan dan organisasi [$t(703.748) = 1.362$, $p = .174 > .05$], dan kaedah analitik

serta aspek penggunaan bahasa dan mekanis [$t(685.091) = 1.173$, $p = .241 > .05$] (Jadual 4.4 dan Jadual 4.6) adalah tidak signifikan.

Jadual 4.3

Keputusan Ujian t Bagi Tugasan Karangan Berlainan Berdasarkan Pemarkahan Kaedah Holistik Serta Aspek Kandungan Dan Organisasi

Tugasan Karangan	*M*	*SD*	*t*	*df*	*p*
Naratif*	6.60	1.424	6.097	670.480	.000**
Pendedahan*	5.85	1.871			

Nota. *N=360; **Signifikan pada aras keertian .05.

Jadual 4.4

Keputusan Ujian t Bagi Tugasan Karangan Berlainan Berdasarkan Pemarkahan Kaedah Analitik Serta Aspek Kandungan Dan Organisasi

Tugasan Karangan	*M*	*SD*	*t*	*df*	*p*
Naratif*	6.62	1.495	1.362	703.748	.174
Pendedahan*	6.46	1.726			

Nota. *N=360; Tidak signifikan pada aras keertian .05.

Jadual 4.5

Keputusan Ujian t Bagi Tugasan Karangan Berlainan Berdasarkan Pemarkahan Kaedah Holistik Serta Aspek Penggunaan Bahasa Dan Mekanis

Tugasan Karangan	*M*	*SD*	*t*	*df*	*p*
Naratif*	6.88	1.493	3.877	694.414	.000**
Pendedahan*	6.40	1.799			

Nota. *N=360; **Signifikan pada aras keertian .05.

Jadual 4.6

Keputusan Ujian t Bagi Tugasan Karangan Berlainan Berdasarkan Pemarkahan Kaedah Analitik Serta Aspek Penggunaan Bahasa Dan Mekanis

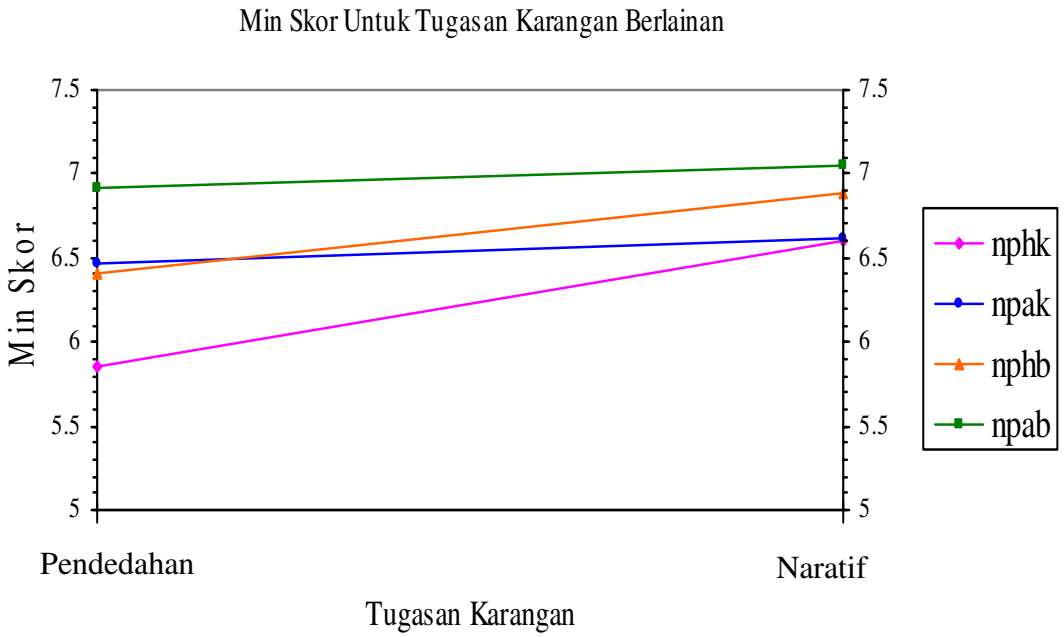
Tugasan Karangan	*M*	*SD*	*t*	*df*	*p*
Naratif*	7.05	1.375	1.173	685.091	.241
Pendedahan*	6.91	1.719			

Nota. *N=360; Tidak signifikan pada aras keertian .05.

Selain itu, keputusan ujian *Levene* untuk kesetaraan varian menunjukkan perbezaan sisihan piawai untuk karangan berunsur naratif dan karangan berunsur pendedahan berdasarkan prosedur pemarkahan yang berlainan adalah berbeza secara signifikan. Ini menunjukkan varian bagi kedua-dua tugasan karangan adalah tidak homogen. Memandangkan tugasan karangan berlainan dalam semua prosedur pemarkahan adalah signifikan berdasarkan ujian *Levene*, dan sisihan piawai bagi karangan berunsur pendedahan adalah lebih tinggi berbanding dengan karangan berunsur naratif dalam semua prosedur pemarkahan yang terlibat, maka ini mungkin menerangkan bahawa ralat pemeriksa bagi pemarkahan karangan berunsur pendedahan adalah lebih besar manakala ralat pemarkahan karangan berunsur naratif secara relatif adalah lebih kecil. Ini selanjutnya menerangkan bahawa tugasan karangan berlainan, pemarkahan pemeriksa, kaedah pemarkahan dan aspek pemarkahan berkemungkinan mengakibatkan keberubahan skor.

Semua prosedur pemarkahan dalam kajian ini juga menunjukkan min skor bagi karangan berunsur pendedahan adalah lebih rendah daripada karangan berunsur naratif (lihat Jadual 4.2 dan Rajah 4.1). Antaranya, skor calon bagi karangan berunsur pendedahan dan karangan berunsur naratif adalah paling hampir antara satu

sama lain berdasarkan prosedur pemarkahan kaedah analitik serta aspek penggunaan bahasa dan mekanis. Jika diteliti, kesan saiz bagi kedua-dua buah tugasan karangan adalah paling kecil iaitu $\eta^2 = .002$ dengan menggunakan prosedur pemarkahan tersebut. Selain itu, data statistik yang diperoleh dalam Jadual 4.2 juga menunjukkan aras kesukaran dan ralat pemarkahan bagi karangan berunsur pendedahan adalah lebih tinggi daripada karangan berunsur naratif. Dengan itu, adalah berasas untuk menganggap bahawa pengaruh faktor tugasan karangan terhadap prosedur pemarkahan kaedah analitik serta aspek penggunaan bahasa dan mekanis mungkin lebih kecil berbanding dengan tiga prosedur pemarkahan yang lain.



Rajah 4.1. Hubungan min skor untuk tugasan karangan berlainan berdasarkan prosedur pemarkahan masing-masing.

Nota. nphk = karangan naratif dan pendedahan berdasarkan kaedah holistik serta aspek kandungan dan organisasi; npak = karangan naratif dan pendedahan berdasarkan kaedah analitik serta aspek kandungan dan organisasi; nphb = karangan naratif dan pendedahan berdasarkan kaedah holistik serta aspek penggunaan bahasa dan mekanis; npab = karangan naratif dan pendedahan berdasarkan kaedah analitik serta aspek penggunaan bahasa dan mekanis.

Secara kesimpulan, berdasarkan analisis statistik deskriptif dan statistik inferensi di atas, soalan kajian pertama iaitu terdapat kemungkinan bahawa faktor tugasan karangan berlainan, pemeriksa, aspek pemarkahan dan kaedah pemarkahan mempengaruhi keberubahan skor adalah terjawab.

4.1.3 Analisis keputusan komponen-komponen varian dalam kajian G

Kajian G dalam kajian ini menggunakan reka bentuk separa tersarang dua faset model rawak. Setiap calon (p) menjawab semua tugasan (t) dan dinilai oleh kumpulan pemeriksa (r) yang berlainan iaitu pemeriksa tersarang dalam tugasan sementara pemeriksa dan tugasan adalah tersilang dengan calon. Ia boleh dilambangkan dengan $p \times (r : t)$. Semua varian dalam kajian ini diandaikan rawak.

Jadual 4.7 hingga Jadual 4.10 masing-masing memaparkan sumber variasi, darjah kebebasan, jumlah kuasa dua, min kuasa dua, anggaran komponen varian dan peratusan setiap komponen varian bagi keputusan reka bentuk $p \times (r : t)$ kajian G berdasarkan empat prosedur pemarkahan yang berlainan. Keputusan-keputusan kajian G untuk prosedur pemarkahan yang berlainan telah ditunjukkan dalam Lampiran S 1 hingga Lampiran S 4. Analisis tersebut adalah berdasarkan tugasan karangan berlainan yang dinilai oleh kaedah pemarkahan dan aspek pemarkahan yang berlainan. Berdasarkan reka bentuk $p \times (r : t)$ separa tersarang dua faset ini, setiap prosedur pemarkahan mempunyai lima komponen varian. Komponen-komponen varian tersebut adalah komponen varian yang berkaitan dengan:

(a) calon [$\sigma^2(p)$],

(b) tugasan [$\sigma^2(t)$],

(c) pemeriksa tersarang dalam tugasan [$\sigma^2(r:t)$],

(d) interaksi calon × tugasan [$\sigma^2(pt)$], dan

(e) interaksi calon × pemeriksa (tersarang dalam tugasan) serta baki ralat yang tidak dapat dileraikan [$\sigma^2(pr{:}t,e)$] atau komponen reja.

Semua anggaran komponen varian dalam kajian G adalah berdasarkan cerapan tunggal iaitu satu calon, satu tugasan dan satu pemeriksa.

Seperti yang ditunjukkan dalam Jadual 4.7 hingga Jadual 4.10, analisis-analisis berasingan tersebut menghasilkan pola yang agak sama dari segi anggaran komponen varian. Ini dapat dilihat menerusi taburan amaun varian tertentu. Sebahagian besar daripada amaun varian tertumpu kepada tiga komponen varian sahaja iaitu $\sigma^2(p)$, $\sigma^2(pt)$ dan $\sigma^2(pr{:}t,e)$ dengan komponen varian untuk calon $\sigma^2(p)$ merupakan komponen yang terbesar merentas semua prosedur pemarkahan. Ketiga-tiga komponen tersebut meliputi lebih daripada 90% bagi varian keseluruhan dalam setiap prosedur pemarkahan. Walaupun komponen varian $\sigma^2(p)$ merupakan komponen paling besar, namun peratusan amaun varian yang diperolehnya adalah agak berbeza merentas empat prosedur pemarkahan yang berlainan (daripada 49.7% hingga 65.0% bagi varian keseluruhan untuk setiap prosedur pemarkahan). Ini menunjukkan calon adalah berbeza secara sistematik dalam skor. Secara perbandingan, prosedur pemarkahan bagi kaedah analitik serta aspek penggunaan bahasa dan mekanis memperoleh peratusan amaun varian yang paling besar iaitu 65.0% (Jadual 4.10) manakala kaedah holistik serta aspek kandungan dan organisasi adalah paling kecil iaitu 49.7% (Jadual 4.7). Komponen varian untuk calon [$\sigma^2(p)$] ini juga merupakan varian skor semesta atau varian skor benar dalam operasi kajian D yang dijalankan seterusnya.

Jadual 4.7

Anggaran Komponen-Komponen Varian Bagi Karangan Berbeza Bentuk Berdasarkan Kaedah Holistik Serta Aspek Kandungan Dan Organisasi Untuk Reka Bentuk p ×(r: t) Kajian G

Sumber Variasi	*df*	*SS*	*MS*	Anggaran Komponen Varian	Peratus Jumlah Varian (%)
p	119	1421.73	11.95	1.534	49.7
t	1	105.80	105.80	0.280	9.1
r:t	4	11.04	2.76	0.019	0.6
pt	119	326.20	2.74	0.745	24.2
pr:t,e	476	240.96	0.51	0.506	16.4

Jadual 4.8

Anggaran Komponen-Komponen Varian Bagi Karangan Berbeza Bentuk Berdasarkan Kaedah Analitik Serta Aspek Kandungan Dan Organisasi Untuk Reka Bentuk p ×(r: t) Kajian G

Sumber Variasi	*df*	*SS*	*MS*	Anggaran Komponen Varian	Peratus Jumlah Varian (%)
p	119	1394.49	11.72	1.571	59.8
t	1	4.84	4.84	0.006	0.2
r:t	4	2.94	0.74	0.003	0.1
pt	119	273.00	2.29	0.623	23.7
pr:t,e	476	201.72	0.42	0.424	16.1

Jadual 4.9

Anggaran Komponen-Komponen Varian Bagi Karangan Berbeza Bentuk Berdasarkan Kaedah Holistik Serta Aspek Penggunaan Bahasa Dan Mekanis Untuk Reka Bentuk p ×(r: t) Kajian G

Sumber Variasi	*df*	*SS*	*MS*	Anggaran Komponen Varian	Peratus Jumlah Varian (%)
p	119	1477.89	12.42	1.698	59.5
t	1	40.61	40.61	0.097	3.4
r:t	4	15.17	3.79	0.028	1.0
pt	119	265.89	2.23	0.601	21.0
pr:t,e	476	204.83	0.43	0.430	15.1

Jadual 4.10

Anggaran Komponen-Komponen Varian Bagi Karangan Berbeza Bentuk Berdasarkan Kaedah Analitik Serta Aspek Penggunaan Bahasa Dan Mekanis Untuk Reka Bentuk p ×(r: t) Kajian G

Sumber Variasi	*df*	*SS*	*MS*	Anggaran Komponen Varian	Peratus Jumlah Varian (%)
p	119	1322.47	11.11	1.593	65.0
t	1	3.20	3.20	0.000	0.0
r:t	4	23.97	5.99	0.046	1.9
pt	119	184.80	1.55	0.371	15.1
pr:t,e	476	209.37	0.44	0.440	18.0

Dari segi anggaran peratusan varian merentas prosedur pemarkahan yang berlainan, komponen varian untuk $\sigma^2(pt)$ merupakan komponen kedua terbesar (daripada 15.1% hingga 24.2%) selepas komponen varian $\sigma^2(p)$ kecuali prosedur kaedah analitik serta aspek penggunaan bahasa dan mekanis. Dalam prosedur tersebut, $\sigma^2(pt)$ merupakan komponen varian ketiga terbesar (15.1%) selepas komponen reja $\sigma^2(pr{:}t,e)$ (18.0%). Berdasarkan taburan varian $\sigma^2(pt)$, prosedur kaedah holistik serta aspek kandungan dan organisasi memiliki peratusan amaun varian yang paling besar (24.2%) manakala amaun varian bagi prosedur kaedah analitik serta aspek penggunaan bahasa dan mekanis adalah paling kecil (15.1%). Komponen varian yang besar untuk interaksi calon × tugasan ini menunjukkan bahawa kedudukan relatif bagi skor calon adalah berlainan berdasarkan tugasan yang berbeza.

Komponen varian untuk reja $\sigma^2(pr{:}t,e)$ merupakan sumber variasi yang agak besar dalam setiap prosedur pemarkahan (daripada 15.1% hingga 18.0%). Ini menunjukkan bahawa sebahagian daripada varian dalam prosedur-prosedur pemarkahan ini adalah disebabkan oleh: (a) perbezaan yang agak besar dalam kedudukan relatif calon daripada satu pemeriksa kepada pemeriksa lain [interaksi calon × pemeriksa (tersarang dalam tugasan) yang agak besar], (b) sumber-sumber varian yang tidak dapat diukur dalam kajian ini (ralat sistematik dan / atau rawak) adalah agak besar (interaksi *prt, e*), atau (c) kedua-duanya sekali.

Julat komponen varian untuk tugasan $\sigma^2(t)$ adalah agak besar dalam prosedur pemarkahan yang berlainan (daripada 0% hingga 9.1%). Komponen varian $\sigma^2(t)$ mewakili bahagian varian yang kecil sahaja dalam semua prosedur pemarkahan

kecuali prosedur kaedah holistik serta aspek kandungan dan organisasi (9.1%). Manakala kaedah analitik serta aspek penggunaan bahasa dan mekanis merupakan prosedur pemarkahan yang mempunyai varian paling kecil merentas tugasan (0%). Ini menerangkan bahawa perbezaan dalam aras kesukaran bagi tugasan karangan pada keseluruhannya adalah sangat kecil.

Daripada keempat-empat prosedur pemarkahan tersebut, komponen varian $\sigma^2(r{:}t)$ memperlihat peratusan amaun varian yang paling kecil sekali iaitu daripada 0.1% hingga 1.9%. Ini bermakna kesan pemeriksa tersarang dalam tugasan adalah sangat kecil merentas prosedur pemarkahan yang berlainan. Komponen varian $\sigma^2(r{:}t)$ adalah paling kecil dalam prosedur kaedah analitik serta aspek kandungan dan organisasi (0.1%) dan paling besar dalam prosedur kaedah analitik serta aspek penggunaan bahasa dan mekanis (1.9%). Untuk komponen varian ini, kesan utama bagi pemeriksa [$\sigma^2(r)$] dan interaksi pemeriksa × tugasan [$\sigma^2(rt)$] adalah terbaur (*confounded*).

Demi tujuan kajian untuk membuat keputusan yang melibatkan kedudukan relatif calon atau keputusan relatif berdasarkan prosedur pemarkahan yang berlainan, komponen-komponen varian yang berkaitan dengan objek pengukuran atau calon iaitu $\sigma^2(pt)$ dan $\sigma^2(pr{:}t,e)$ sahaja perlu diambil kira dan dianalisis. Jadual 4.11 menunjukkan peratusan komponen varian dan nilai pekali G bagi cerapan tunggal dalam prosedur pemarkahan berlainan yang dilihat daripada perspektif keputusan relatif. Daripada Jadual 4.11 dan Rajah 4.2, jelas menunjukkan komponen varian bagi calon atau objek pengukuran adalah paling besar dan mengatasi varian ralat

$\sigma^2(pt)$ dan $\sigma^2(pr{:}t,e)$ merentas semua prosedur pemarkahan dalam kajian ini. Nilainya juga adalah berbeza-beza di bawah prosedur pemarkahan yang berlainan.

Jadual 4.11

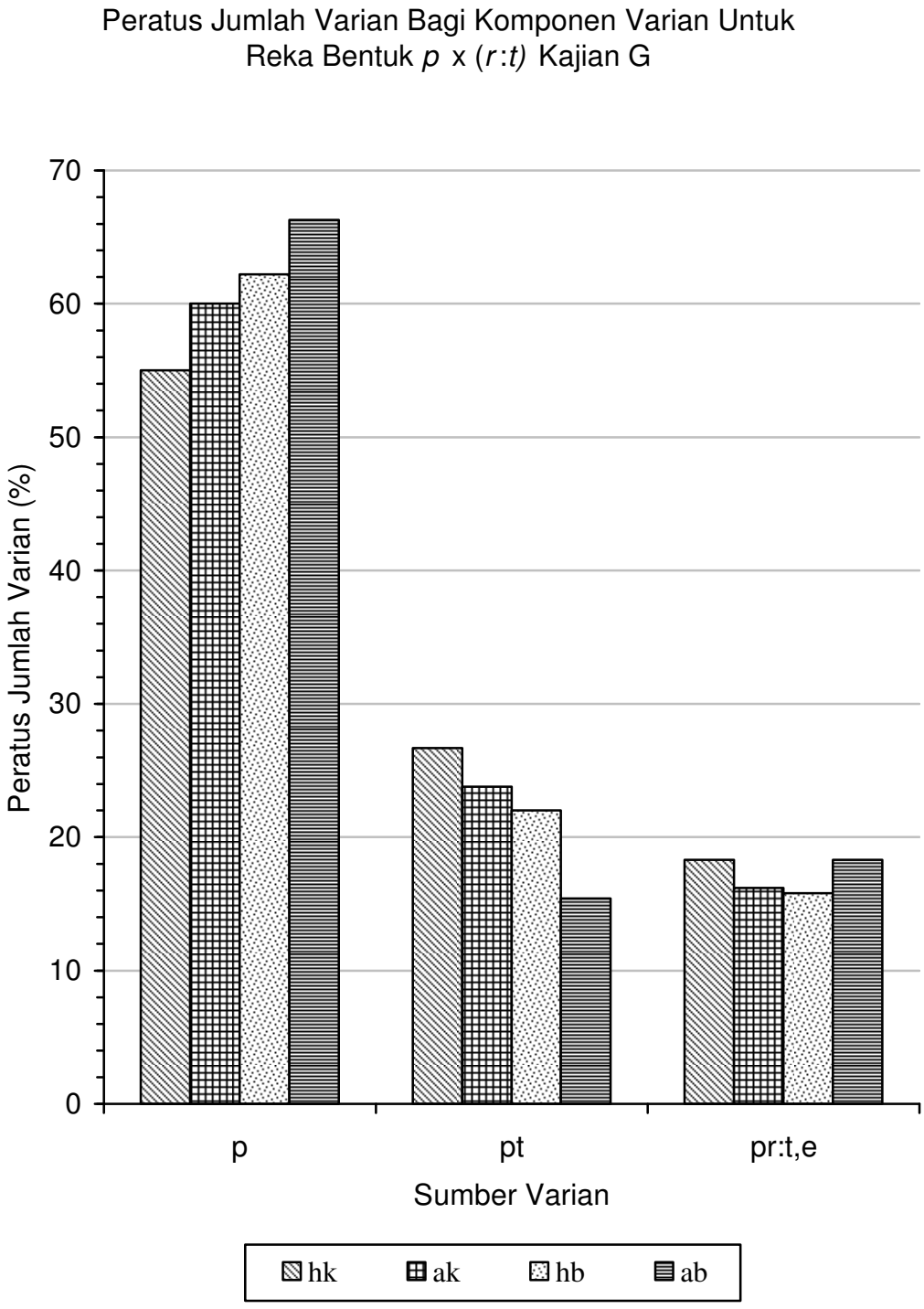
Peratus Komponen Varian Dan Pekali G Bagi Prosedur Pemarkahan Yang Berlainan Berdasarkan Keputusan Relatif

	Kaedah Holistik Serta Aspek Kandungan Dan Organisasi		Kaedah Analitik Serta Aspek Kandungan Dan Organisasi		Kaedah Holistik Serta Aspek Penggunaan Bhs. Dan Mekanis		Kaedah Analitik Serta Aspek Penggunaan Bhs. Dan Mekanis	
Kesan	σ^2	%	σ^2	%	σ^2	%	σ^2	%
p	1.534	55.0	1.571	60.0	1.698	62.2	1.593	66.3
pt	0.745	26.7	0.623	23.8	0.601	22.0	0.371	15.4
pr:t,e	0.506	18.3	0.424	16.2	0.430	15.8	0.440	18.3
Jumlah	2.789	100.0	2.618	100.0	2.729	100.0	2.404	100.0
$\sigma^2(\delta)$	1.251		1.047		1.031		0.771	
*($E\rho^2$)	0.55		0.60		0.62		0.66	

Nota. $\sigma^2(\delta)$ = varian ralat relatif; $E\rho^2$= pekali G.

*Anggaran pekali G ini adalah berasaskan cerapan tunggal dalam peringkat kajian G.

Bagi prosedur pemarkahan yang menggunakan kaedah holistik serta aspek kandungan dan organisasi, nilai $\sigma^2(p)$ merupakan kira-kira 55% daripada varian keseluruhan dan adalah lebih kurang sekali ganda nilai $\sigma^2(pt)$ dan dua kali ganda nilai $\sigma^2(pr{:}t,e)$ (lihat Rajah 4.2). Secara perbandingan, prosedur pemarkahan ini adalah paling kurang memuaskan di antara semua prosedur pemarkahan yang digunakan memandangkan amaun varian benar iaitu $\sigma^2(p)$ adalah terendah sedangkan varian benar merupakan varian yang hendak diukur. Ini juga bermakna ralat pengukuran bagi prosedur pemarkahan ini adalah paling besar kerana ketepatannya untuk mentaksir kebolehan calon dalam kemahiran menulis secara relatif adalah rendah.



Rajah 4.2. Peratus jumlah varian bagi komponen-komponen varian untuk reka bentuk $p \times (r{:}t)$ dalam kajian G bagi empat prosedur pemarkahan yang berlainan berdasarkan keputusan relatif.

Nota. p = calon; pt = interaksi calon dan tugasan; pr:t, e = interaksi calon dan pemeriksa (tersarang dalam tugasan) serta baki ralat; hk = kaedah holistik serta aspek kandungan dan organisasi; ak = kaedah analitik serta aspek kandungan dan organisasi; hb = kaedah holistik serta aspek penggunaan bahasa dan mekanis; ab = kaedah analitik serta aspek penggunaan bahasa dan mekanis.

Namun begitu, kesan calon yang terdapat dalam prosedur pemarkahan kaedah analitik serta aspek kandungan dan organisasi pula adalah lebih baik daripada kaedah holistik walaupun menggunakan aspek pemarkahan yang sama, iaitu nilai *p* adalah kira-kira 60% daripada jumlah varian. Begitu juga keadaan bagi aspek penggunaan bahasa dan mekanis, kesan calon dalam kaedah analitik (66.3%) adalah lebih baik daripada kaedah holistik (62.2%).

Secara teliti, komponen varian bagi calon yang menggunakan kaedah analitik serta aspek penggunaan bahasa dan mekanis adalah paling besar iaitu 1.593 (66.3%) berbanding dengan semua prosedur pemarkahan yang lain dalam kajian ini. Ini bermakna kesan prosedur pemarkahan ini adalah paling baik walaupun nilai varian benar bagi objek pengukuran masih kurang daripada keadaan yang ideal. Namun begitu, komponen varian bagi calon yang menggunakan kaedah holistik dengan aspek pemarkahan yang sama telah memaparkan nilai yang secara relatif adalah lebih kecil iaitu kira-kira 62.2% (1.698) daripada jumlah varian.

Sekiranya melihat nilai pekali G untuk keputusan relatif berdasarkan prosedur pemarkahan yang berlainan (Jadual 4.11 dan Rajah 4.2), boleh dikatakan pada keseluruhannya objek pengukuran bagi keempat-empat prosedur pemarkahan dalam kajian ini adalah sederhana tinggi dan masih kurang memuaskan. Kaedah analitik serta aspek penggunaan bahasa dan mekanis mempunyai nilai *p* yang paling besar iaitu 66.3% daripada jumlah varian, hanya sekadar dua pertiga daripada varian benar yang dikehendaki. Ini juga bermakna masih terdapat faktor-faktor lain yang boleh mempengaruhi varian skor karangan walaupun objek pengukuran merupakan dominan utama bagi varian skor karangan. Namun begitu, kesan keempat-empat

prosedur pemarkahan adalah berbeza dari segi kebolehan generalisasi. Jika diteliti, kebolehan generalisasi bagi aspek pemarkahan dan kaedah pemarkahan yang berlainan adalah berbeza. Keadaan ini dapat dilihat berdasarkan nilai pekali G yang diperoleh menerusi cerapan tunggal dalam reka bentuk kajian G bagi kaedah pemarkahan dan aspek pemarkahan yang berlainan (Jadual 4.11). Jika ditinjau dari segi aspek pemarkahan, kebolehan generalisasi bagi aspek penggunaan bahasa dan mekanis (.62 dan .66) pada keseluruhannya adalah lebih kuat daripada aspek kandungan dan organisasi (.55 dan .60). Manakala dilihat dari segi kaedah pemarkahan pula, kebolehan generalisasi bagi kaedah analitik (.60 dan .66) secara relatif adalah lebih baik daripada kaedah holistik (.55 dan .62).

Untuk menganalisis kesan tugasan berdasarkan prosedur pemarkahan yang berlainan, maka anggaran dan peratusan komponen varian bagi varian interaksi calon dan tugasan $\sigma^2(pt)$ harus ditinjau. Dari segi aspek kandungan dan organisasi (berdasarkan kaedah pemarkahan yang berlainan), varian $\sigma^2(pt)$ mencatatkan 0.745 (26.7%) dan 0.623 (23.8%) masing-masing. Amaun varian yang meliputi kira-kira suku daripada jumlah varian ini adalah agak besar dan merupakan komponen varian kedua terbesar selain objek pengukuran. Dalam aspek kandungan dan organisasi serta kaedah holistik, varian $\sigma^2(pt)$ adalah hampir separa daripada objek pengukuran dan amaun ini juga adalah paling besar merentas semua prosedur pemarkahan. Ini menunjukkan bahawa varian skor bagi pentaksiran yang menggunakan aspek kandungan dan organisasi adalah besar. Bagi aspek penggunaan bahasa dan mekanis (berdasarkan kaedah pemarkahan yang berlainan) pula, varian $\sigma^2(pt)$ adalah 0.601 (22.0%) dan 0.371 (15.4%) masing-masing. Walaupun amaun variannya tidak sebesar aspek kandungan dan organisasi tetapi mereka masih merupakan komponen

varian kedua terbesar selepas objek pengukuran.

Sementara itu, jika ditinjau dari segi kaedah pemarkahan (berdasarkan aspek pemarkahan yang berlainan), kesan faktor tugasan (*pt*) yang ditunjukkan dalam kaedah holistik secara relatif adalah lebih besar (26.7% dan 22.0%) daripada kaedah analitik (23.8% dan 15.4%). Secara keseluruhan, faktor kesan karangan berbeza tugasan adalah wujud dalam tugasan karangan berdasarkan kajian ini. Malah sumbangan varian skor bagi kesan tugasan juga adalah berbeza berdasarkan prosedur pemarkahan yang berlainan. Perbezaan ini adalah lebih jelas terutamanya bagi aspek kandungan dan organisasi. Dalam aspek tersebut, varian atau ralat skor adalah lebih besar tanpa mengira sama ada kaedah holistik atau kaedah analitik.

Penganalisisan mengenai kesan tentang faktor pemeriksa adalah rumit sedikit. Memandangkan kajian ini menggunakan reka bentuk separa tersarang iaitu pemeriksa tersarang dalam tugasan (*r:t*), oleh itu kesan pemeriksa adalah dibatasi oleh tugasan karangan yang berlainan. Dengan itu, kesan pemeriksa adalah terbaur atau dicemari oleh tugasan karangan yang berlainan, maka pentafsiran ke atasnya adalah kurang tepat. Bagaimanapun, daripada dapatan kajian yang diperoleh, kesan pemeriksa masih mempunyai trend tertentu yang boleh dianalisis. Jika dilihat dari aspek kandungan dan organisasi berdasarkan kaedah pemarkahan yang berbeza, nilai komponen varian $\sigma^2(r{:}t)$ ialah 0.019 (kaedah holistik) dan 0.003 (kaedah analitik) masing-masing. Manakala bagi aspek penggunaan bahasa dan mekanis pula, nilai komponen varian $\sigma^2(r{:}t)$ ialah 0.028 (kaedah holistik) dan 0.046 (kaedah analitik) (lihat Jadual 4.9 dan Jadual 4.10). Oleh itu, untuk menganalisis kesan pemeriksa mengikut aspek pemarkahan, kita boleh berpandukan maklumat tersebut di samping

mengambil kira nilai varian dan peratusan komponen varian bagi faktor *pr:t,e*. Melalui analisis dan perbandingan, adalah jelas bahawa varian ralat pemeriksa bagi aspek penggunaan bahasa dan mekanis pada keseluruhannya adalah lebih besar daripada aspek kandungan dan organisasi.

Untuk menganalisis kesan pemeriksa mengikut kaedah pemarkahan pula, jika diteliti dapatan yang diperoleh, keputusan perbezaan varian ralat pemeriksa bagi pentaksiran kaedah holistik (0.019 dan 0.028) dan kaedah analitik (0.003 dan 0.046) adalah bercampur-campur. Oleh itu adalah sukar untuk menentukan sama ada kaedah holistik atau kaedah analitik memberi pengaruh yang lebih besar kepada kesan pemeriksa. Meskipun begitu, jika ditinjau bersama dengan faktor *pr:t,e*, maka varian ralat pemeriksa dalam kaedah holistik adalah lebih besar sedikit daripada kaedah analitik walaupun prosedur kaedah analitik serta aspek penggunaan bahasa dan mekanis memaparkan komponen varian $\sigma^2(r{:}t)$ yang paling besar (0.046). Namun begitu, trend ini adalah tidak begitu jelas.

Di samping itu, analisis komponen-komponen varian yang berkaitan dengan tugasan karangan dan pemeriksa di atas juga menunjukkan bahawa kesan tugasan yang berlainan dan kesan pemeriksa adalah bersandar kepada prosedur pemarkahan yang berlainan.

Dengan ini, maka terjawablah soalan kajian kedua iaitu sejauh manakah perkaitan antara kesan tugasan karangan berlainan dan kesan pemeriksa terhadap prosedur pemarkahan yang berlainan dari segi kebergantungan skor karangan berdasarkan kerangka kajian G.

4.2 Analisis Keputusan Kajian D

4.2.1 Analisis keputusan reka bentuk separa tersarang $p \times (R:T)$ model rawak

Jadual 4.12 hingga Jadual 4.15 dan Rajah 4.3 hingga Rajah 4.10 memaparkan anggaran pekali G ($E\rho^2$) menerusi operasi pelbagai kajian D untuk reka bentuk $p \times (R:T)$ separa tersarang model rawak bagi bilangan tugasan dan bilangan pemeriksa daripada satu hingga lapan berdasarkan empat prosedur pemarkahan yang berlainan. Menurut Nunnally dan Burnstein (1994), nilai pekali kebolehpercayaan .90 adalah sesuai untuk tujuan membuat keputusan tentang individu (lihat juga Linn dan Gronlund, 2005). Oleh itu, pekali kebolehpercayaan .90 dianggap sebagai kriteria bagi pekali kebolehpercayaan yang diingini untuk membuat keputusan relatif dan menganalisis kebergantungan skor prosedur pemarkahan yang berlainan dalam kajian ini. Keputusan-keputusan kajian D untuk prosedur pemarkahan yang berlainan telah ditunjukkan dalam Lampiran T 1 hingga Lampiran T 4. Analisis-analisis berikut adalah berdasarkan prosedur pemarkahan yang berlainan secara berasingan.

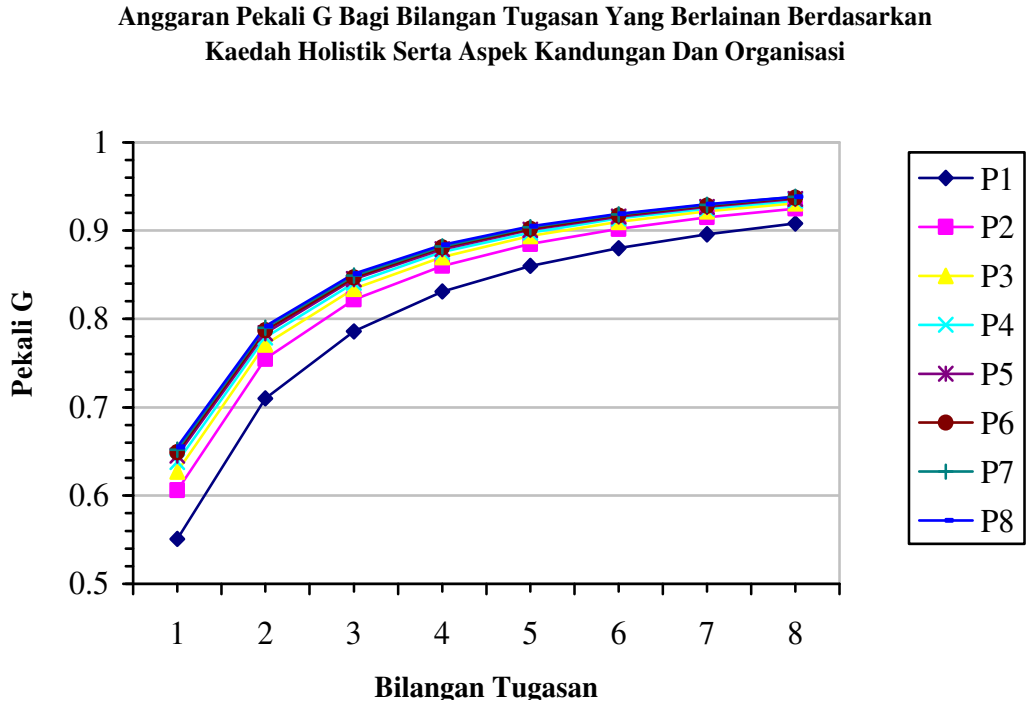
Jadual 4.12 serta Rajah 4.3 dan Rajah 4.4 memaparkan anggaran pekali G ($E\rho^2$) dalam kajian D untuk reka bentuk separa tersarang $p \times (R: T)$ model rawak bagi bilangan tugasan dan bilangan pemeriksa yang berlainan berdasarkan kaedah holistik serta aspek kandungan dan organisasi. Seperti yang ditunjukkan dalam Rajah 4.3, berdasarkan bilangan tugasan satu hingga lapan, pertambahan bilangan tugasan dengan menggunakan kaedah holistik serta aspek kandungan dan organisasi secara relatif mempunyai impak yang lebih besar terhadap kebergantungan skor. Sebaliknya, impak bagi pertambahan bilangan pemeriksa berdasarkan prosedur pemarkahan yang sama terhadap kebergantungan skor secara relatif adalah lebih kecil (Rajah 4.4).

Jadual 4.12

Anggaran Pekali G ($E\rho^2$) Bagi Bilangan Tugasan Dan Bilangan Pemeriksa Yang Berlainan Berdasarkan Kaedah Holistik Serta Aspek Kandungan Dan Organisasi

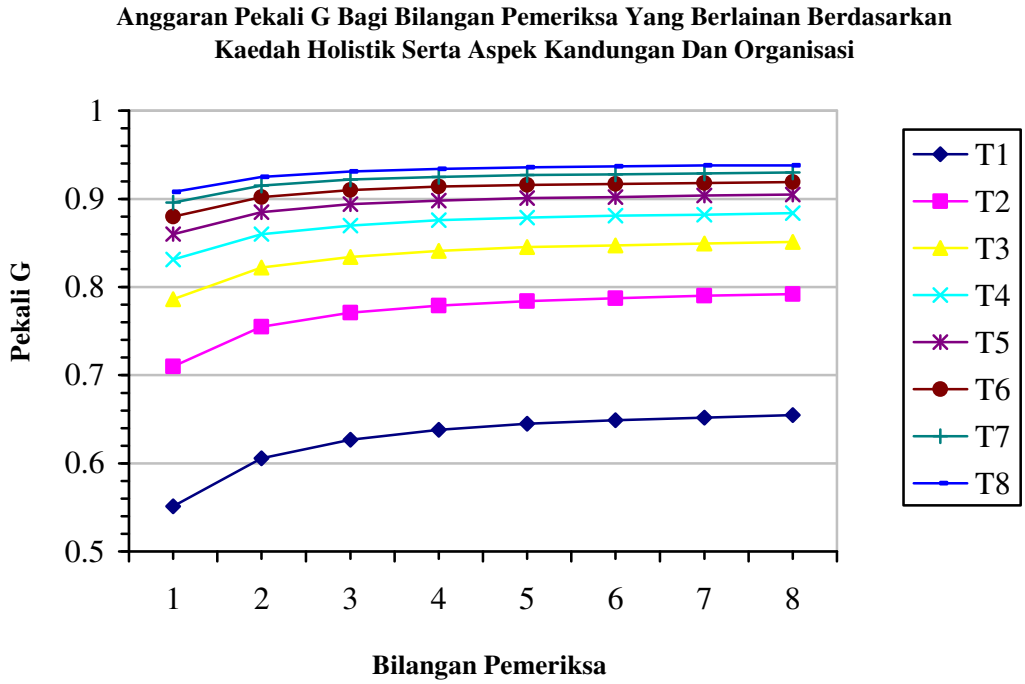
Bilangan Pemeriksa ($\acute{n}_r$)	Bilangan Tugasan ($\acute{n}_t$)							
	1	2	3	4	5	6	7	8
1	.551	.710	.786	.831	.860	.880	.896	.908
2	.606	.755	.822	.860	.885	.902	.915	.925
3	.627	.771	.834	.870	.894	.910	.922	.931
4	.638	.779	.841	.876	.898	.914	.925	.934
5	.645	.784	.845	.879	.901	.916	.927	.936
6	.649	.787	.847	.881	.902	.917	.928	.937
7	.652	.790	.849	.882	.904	.918	.929	.938
8	.655	.792	.851	.884	.905	.919	.930	.938

Nota. Berdasarkan data asal (n_p = *120*, n_t = *2*, n_r = *3*) dengan reka bentuk separa tersarang $p \times (r{:}t)$.



Rajah 4.3. Pekali G bagi bilangan tugasan yang berlainan berdasarkan kaedah holistik serta aspek kandungan dan organisasi.

Nota. P1= pemeriksa pertama; P2 = pemeriksa kedua dan seterusnya.



Rajah 4.4. Pekali G bagi bilangan pemeriksa yang berlainan berdasarkan kaedah holistik serta aspek kandungan dan organisasi.

Nota. T1= tugasan pertama; T2 = tugasan kedua dan seterusnya.

Untuk meninjau impak bilangan tugasan karangan terhadap kebergantungan skor berdasarkan prosedur pemarkahan ini, kita boleh meneliti perubahan bilangan tugasan karangan (Jadual 4.12 dan Rajah 4.3). Berdasarkan senario pemeriksa tunggal, apabila bilangan tugasan bertambah daripada satu hingga tiga, pekali G meningkat secara drastik iaitu daripada .551 kepada .786 iaitu pertambahan sebanyak .235 atau 42.7%. Secara spesifik, pekali G bertambah sebanyak .159 (.710–.551) atau 28.9% untuk dua tugasan pertama dan pertambahan pekali G sebanyak .086 (.786 –.710) untuk bilangan tugasan ketiga. Bagaimanapun, pertambahan bilangan tugasan yang selanjutnya daripada empat hingga lapan iaitu pertambahan sebanyak lima tugasan, hanya meningkatkan pekali G sebanyak .077 (.908 –.831) atau 9.3%. Ini menunjukkan bahawa terdapat pulangan menurun (*diminishing returns*) bagi pekali G apabila lebih banyak tugasan digunakan.

Pola peningkatan pekali G yang lebih kurang sama boleh didapati tanpa mengira bilangan pemeriksa yang digunakan iaitu peningkatan pekali G yang lebih besar akan dicapai apabila bilangan tugasan berubah daripada satu menjadi dua dan pertambahan tugasan selepas itu menampakkan peningkatan pekali G yang menurun (Rajah 4.3). Misalnya apabila $\acute{n}_r = 3$, dan $\acute{n}_t = 1$ bertambah menjadi $\acute{n}_t = 2$, pekali G meningkat sebanyak .144 (.771 – .627) atau 23% manakala pertambahan $\acute{n}_t = 3$ hingga $\acute{n}_t = 8$, hanya meningkatkan pekali G sebanyak .097 (.931 – .834) atau 11.6%; apabila $\acute{n}_r = 8$, dan $\acute{n}_t = 1$ bertambah menjadi $\acute{n}_t = 2$, pekali G meningkat sebanyak .137 (.792–.655) atau 20.9% manakala pertambahan $\acute{n}_t = 3$ hingga $\acute{n}_t = 8$, pekali G hanya meningkat sebanyak .087 (.938 – .851) atau 10.2%.

Secara perbandingan, untuk melihat impak bilangan pemeriksa terhadap kebergantungan skor, kita boleh meneliti perubahan bilangan pemeriksa (Jadual 4.12 dan Rajah 4.4). Berdasarkan senario tugasan tunggal, pekali G meningkat sebanyak .055 (.606–.551) atau 10.0% apabila bilangan pemeriksa bertambah daripada satu kepada dua. Pertambahan bilangan pemeriksa seterusnya daripada tiga hingga lapan hanya menambahkan pekali G sebanyak .028 (.655 – .627) atau 4.5%. Pola peningkatan pekali G yang lebih kurang sama boleh didapati tanpa mengira senario bilangan tugasan yang digunakan. Peningkatan pekali G adalah lebih besar untuk bilangan pemeriksa satu kepada dua namun pertambahan pemeriksa selepas itu menyebabkan pekali G mengalami pulangan menurun. Ini jelas dapat dilihat dalam Rajah 4.4 iaitu peningkatan pekali G bagi pemeriksa akan berkurangan apabila semakin banyak pemeriksa digunakan terutamanya penggunaan bilangan pemeriksa yang keempat dan ke atas tidak banyak memperbaiki kepersisan pengukuran. Bagaimanapun, impak bilangan pemeriksa terhadap pekali G secara relatif adalah

kecil berbanding dengan pertambahan bilangan tugasan.

Penelitian perubahan pekali G berdasarkan kombinasi bilangan tugasan dan pemeriksa, jelas menunjukkan bahawa pertambahan bilangan tugasan nampaknya mempunyai impak yang lebih besar daripada pertambahan bilangan pemeriksa terhadap kebergantungan skor. Misalnya, dalam prosedur pemarkahan ini, pekali G berdasarkan senario tugasan dan pemeriksa tunggal ialah .551. Apabila kombinasi $\acute{n}_t$ = 1, $\acute{n}_r$ = 2 digunakan, pekali G meningkat kepada .606 atau 10.0%. Namun, apabila kombinasi $\acute{n}_t$ = 2, $\acute{n}_r$ = 1 digunakan, pekali G meningkat lebih tinggi lagi iaitu .710 atau 28.9%. Pola yang sama boleh dilihat untuk kombinasi $\acute{n}_t$ = 3, $\acute{n}_r$ = 4 (.841) berbanding dengan $\acute{n}_t$ = 4, $\acute{n}_r$ = 3 (.870); $\acute{n}_t$ = 4, $\acute{n}_r$ = 5 (.879) berbanding dengan $\acute{n}_t$ = 5, $\acute{n}_r$ = 4 (.898) dan seterusnya. Malah dengan satu tugasan, walaupun lapan pemeriksa digunakan, pekali G hanya mencecah .655. Sebaliknya, berdasarkan satu pemeriksa dan lapan tugasan, pekali dapat mencecah kriteria .90 yang dihajati. Sementara itu, pekali G untuk reka bentuk kajian asal bagi prosedur ini iaitu $\acute{n}_t$ = 2, $\acute{n}_r$ = 3 ialah .771, sekiranya kombinasi $\acute{n}_t$ = 3, $\acute{n}_r$ = 2 digunakan, kepersisan pengukuran boleh diperbaiki lagi kerana pekali G akan meningkat kepada .822. Namun begitu, nilai pekali G tersebut masih belum menepati nilai yang dihajatkan iaitu .90. Berdasarkan prosedur pemarkahan ini, pemadanan optimum bagi bilangan tugasan dan pemeriksa berdasarkan kriteria .90 adalah $\acute{n}_t$ = 5, $\acute{n}_r$ = 4. Selain itu, pemadanan lain yang boleh dipilih adalah $\acute{n}_t$ = 6, $\acute{n}_r$ = 2 atau $\acute{n}_t$ = 7, $\acute{n}_r$ = 1 sejajar dengan pertimbangan kos dan masa. Secara kesimpulan, prosedur ini memaparkan impak pertambahan bilangan tugasan adalah lebih besar daripada pertambahan bilangan pemeriksa dalam mencapai pekali yang tinggi.

Jadual 4.13 serta Rajah 4.5 dan Rajah 4.6 memaparkan anggaran pekali G ($E\rho^2$) menerusi operasi pelbagai kajian D untuk reka bentuk separa tersarang $p\times(R\colon T)$ model rawak bagi bilangan tugasan dan bilangan pemeriksa daripada satu hingga lapan berlandaskan kaedah analitik serta aspek kandungan dan organisasi. Dalam prosedur pemarkahan ini, pertambahan bilangan tugasan secara relatif mempunyai impak yang lebih besar terhadap kebergantungan skor. Manakala impak untuk pertambahan bilangan pemeriksa terhadap kebergantungan skor secara relatif pula adalah kecil. Misalnya, pekali G meningkat secara drastik iaitu daripada .60 kepada .75 atau meningkat sebanyak 25.0% apabila $\acute{n}_t = 1$ berubah menjadi $\acute{n}_t = 2$ berdasarkan senario pemeriksa tunggal. Manakala pekali G hanya meningkat sebanyak .053 (8.8%) apabila $\acute{n}_r = 1$ berubah kepada $\acute{n}_r = 2$ berdasarkan senario tugasan tunggal. Untuk pertambahan bilangan tugasan yang selanjutnya daripada tiga hingga lapan iaitu pertambahan sebanyak enam tugasan, memperlihatkan pekali G meningkatkan sebanyak .105 (.923–.818) atau 12.8%. Namun, untuk pertambahan bilangan pemeriksa yang sama iaitu enam orang berdasarkan tugasan tunggal, pekali G hanya meningkat sebanyak .026 (.699 – .673) atau 3.9%. Ini juga menunjukkan bahawa terdapat pulangan menurun dalam pekali G apabila lebih banyak tugasan atau pemeriksa digunakan.

Pola yang sama boleh didapati apabila senario bilangan tugasan atau bilangan pemerika yang seterusnya digunakan iaitu pekali G akan meningkat dengan cepat pada peringkat awalnya dan berkurangan selepas itu (lihat Rajah 4.5 dan Rajah 4.6). Namun begitu, peratusan peningkatan pekali G bagi bilangan tugasan adalah lebih besar berbanding dengan bilangan pemeriksa untuk semua keadaan pertambahan bilangan pada kadar yang sama. Misalnya, apabila $\acute{n}_r = 2$, dan $\acute{n}_t = 3$ berubah kepada

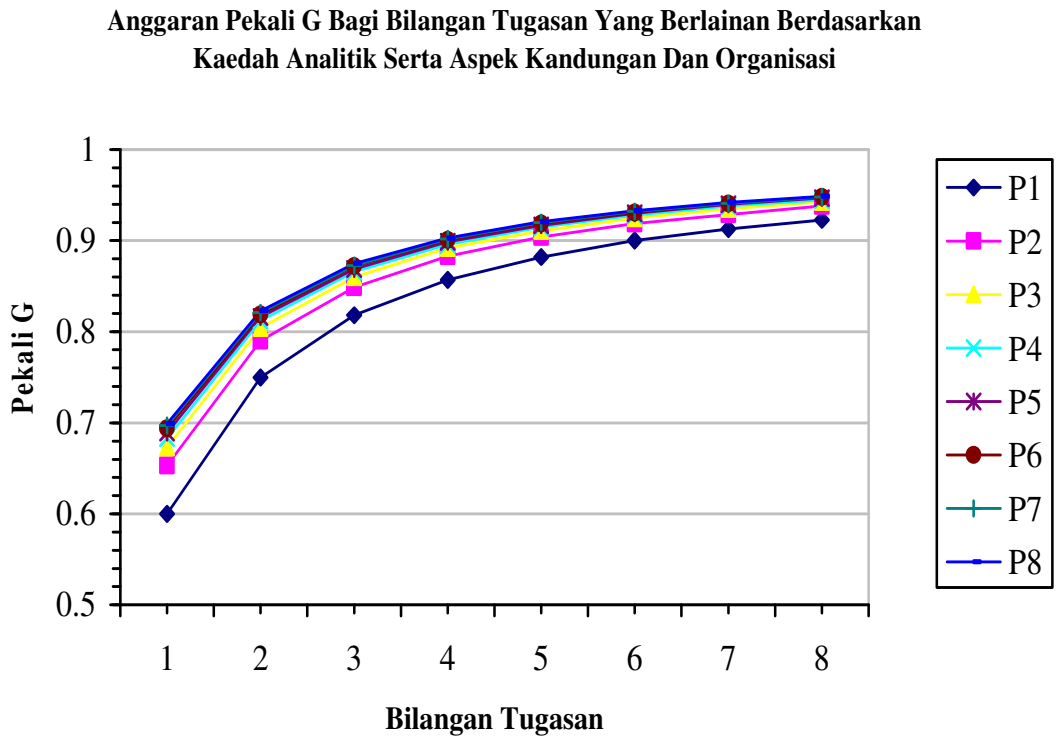
$\acute{n}_t$ = 4, peratusan pekali G untuk pertambahan bilangan tugasan meningkat sebanyak 4.0% (.883 –.849) manakala apabila $\acute{n}_t$ = 2, dan $\acute{n}_r$ = 3 berubah kepada $\acute{n}_r$ = 4, peratusan pekali G untuk pertambahan bilangan pemeriksa hanya meningkat sebanyak 1.0% (.812–.804); apabila $\acute{n}_r$ = 6, dan $\acute{n}_t$ = 3 berubah kepada $\acute{n}_t$ = 4, peratusan pekali G untuk pertambahan bilangan tugasan meningkat sebanyak 3.3% (.901 – .872) manakala apabila $\acute{n}_t$ = 6, dan $\acute{n}_r$ = 3 berubah kepada $\acute{n}_r$ = 4, peningkatan pekali G untuk pertambahan bilangan pemeriksa adalah hanya 0.3%. (.928 – .925).

Jadual 4.13

Keputusan Pekali G Bagi Bilangan Tugasan Dan Bilangan Pemeriksa Yang Berlainan Berdasarkan Kaedah Analitik Serta Aspek Kandungan Dan Organisasi

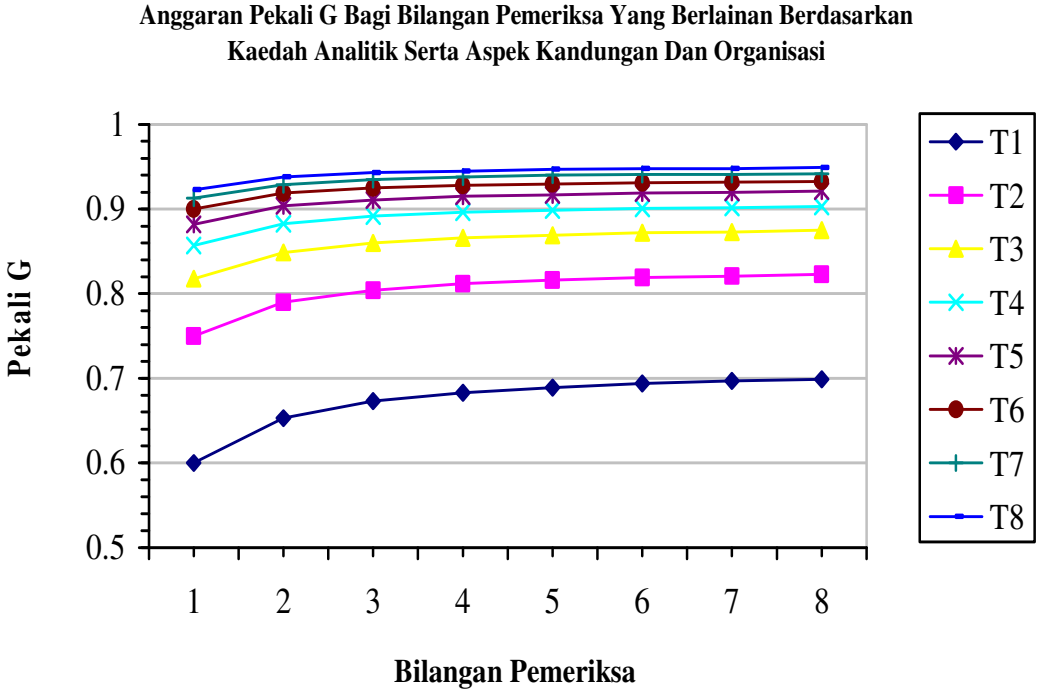
Bilangan Pemeriksa ($\acute{n}_r$)	Bilangan Tugasan ($\acute{n}_t$)							
	1	2	3	4	5	6	7	8
1	.600	.750	.818	.857	.882	.900	.913	.923
2	.653	.790	.849	.883	.904	.919	.929	.938
3	.673	.804	.860	.892	.911	.925	.935	.943
4	.683	.812	.866	.896	.915	.928	.938	.945
5	.689	.816	.869	.899	.917	.930	.940	.947
6	.694	.819	.872	.901	.919	.931	.941	.948
7	.697	.821	.873	.902	.920	.932	.941	.948
8	.699	.823	.875	.903	.921	.933	.942	.949

Nota. Berdasarkan data asal (n_p = *120,* n_t = *2,* n_r = *3*) dengan reka bentuk separa tersarang $p \times (r{:}t)$.



Rajah 4.5. Pekali G bagi bilangan tugasan yang berlainan berdasarkan kaedah analitik serta aspek kandungan dan organisasi.

Nota. P1= pemeriksa pertama; P2 = pemeriksa kedua dan seterusnya.



Rajah 4.6. Pekali G bagi bilangan pemeriksa yang berlainan berdasarkan kaedah analitik serta aspek kandungan dan organisasi.

Nota. T1= tugasan pertama; T2 = tugasan kedua dan seterusnya.

Untuk meninjau perubahan pekali G berasaskan kombinasi bilangan tugasan dan pemeriksa bagi prosedur ini, pertambahan bilangan tugasan nampaknya mempunyai impak yang lebih besar berbanding dengan pertambahan bilangan pemeriksa terhadap kebergantungan skor. Misalnya, berdasarkan senario tugasan dan pemeriksa tunggal, pekali G untuk prosedur pemarkahan ini ialah .60. Apabila kombinasi $\acute{n}_t = 1$, $\acute{n}_r = 2$ digunakan, pekali G meningkat kepada .653 atau 8.8%. Namun, apabila kombinasi ialah $\acute{n}_t = 2$, $\acute{n}_r = 1$, pekali G meningkat lebih tinggi lagi iaitu .75 atau 25%. Pola yang sama boleh dilihat menerusi kombinasi $\acute{n}_t = 3$, $\acute{n}_r = 4$ (.866) berbanding dengan kombinasi $\acute{n}_t = 4$, $\acute{n}_r = 3$ (.892); kombinasi $\acute{n}_t = 4$, $\acute{n}_r = 5$ (.899) berbanding dengan kombinasi $\acute{n}_t = 5$, $\acute{n}_r = 4$ (.915) dan seterusnya yang menunjukkan pertambahan bilangan tugasan memberi kesan yang lebih besar terhadap pekali G berbanding dengan pertambahan bilangan pemeriksa. Malah dengan satu pemeriksa, enam tugasan diperlukan untuk mencecah kriteria .90 yang dihajati. Manakala dengan satu tugasan, walaupun lapan pemeriksa digunakan, pekali G hanya mencecah .70. Sementara itu, pekali G untuk reka bentuk kajian asal dalam prosedur ini iaitu $\acute{n}_t = 2$, $\acute{n}_r = 3$ adalah .80. Sekiranya kombinasi $\acute{n}_t = 3$, $\acute{n}_r = 2$ digunakan, kepersisan pengukuran boleh diperbaiki lagi kerana pekali G akan meningkat kepada .85, walaupun ia masih tidak dapat memenuhi kriteria .90. Untuk mencapai kriteria .90, pemadanan bilangan tugasan dan bilangan pemeriksa yang sama iaitu empat ($\acute{n}_t = 4$, $\acute{n}_r = 4$) adalah diperlukan. Pemadanan lain yang boleh dipertimbangkan adalah $\acute{n}_t = 5$, $\acute{n}_r = 2$ atau $\acute{n}_t = 6$, $\acute{n}_r = 1$.

Jadual 4.14 serta Rajah 4.7 dan Rajah 4.8 memaparkan anggaran pekali G ($E\rho^2$) menerusi operasi pelbagai kajian D untuk reka bentuk separa tersarang $p \times (R : T)$ model rawak bagi bilangan tugasan dan bilangan pemeriksa yang

berlainan berdasarkan kaedah holistik serta aspek penggunaan bahasa dan mekanis. Dalam prosedur pemarkahan ini, pertambahan bilangan tugasan secara relatif mempunyai impak yang lebih besar terhadap kebergantungan skor. Namun, impak untuk pertambahan bilangan pemeriksa terhadap kebergantungan skor secara relatif adalah kecil. Misalnya, berdasarkan senario pemeriksa tunggal, pekali G meningkat secara drastik iaitu daripada .622 kepada .767 atau sebanyak 23.3% apabila $\acute{n}_t = 1$ berubah menjadi $\acute{n}_t = 2$. Manakala berdasarkan senario tugasan tunggal, pekali G hanya meningkat sebanyak .053 (.675 –.622) atau 8.5% apabila $\acute{n}_r = 1$ berubah kepada $\acute{n}_r = 2$. Pertambahan bilangan tugasan yang selanjutnya daripada tiga hingga lapan iaitu pertambahan sebanyak enam tugasan, menampakkan pekali G meningkatkan sebanyak .097 (.929–.832) atau 11.7%. Namun, untuk pertambahan bilangan pemeriksa pada kadar yang sama, pekali G hanya meningkat sebanyak .027 (.722 – .695) atau 3.9%. Ini juga menunjukkan bahawa terdapat pulangan menurun dalam pekali G apabila lebih banyak tugasan atau pemeriksa digunakan.

Pola yang sama boleh didapati apabila bilangan tugasan atau bilangan pemeriksa yang seterusnya digunakan, pekali G sentiasa meningkat dengan cepat pada peringkat awalnya terutamanya untuk pertambahan tugasan atau pemeriksa yang kedua dan selepas itu pekali G akan berkurangan (lihat Rajah 4.7 dan 4.8). Namun begitu, peratusan peningkatan pekali G bagi bilangan tugasan adalah lebih tinggi daripada bilangan pemeriksa untuk semua keadaan pertambahan bilangan pada kadar yang sama. Misalnya, apabila $\acute{n}_r = 2$, dan $\acute{n}_t = 3$ berubah menjadi $\acute{n}_t = 4$, peratusan pekali G untuk bilangan tugasan meningkat sebanyak 3.6% (.893 –.862) manakala apabila $\acute{n}_t = 2$, dan $\acute{n}_r = 3$ berubah kepada $\acute{n}_r = 4$, peratusan pekali G untuk bilangan pemeriksa hanya meningkat sebanyak 0.9% (.827–.820); apabila $\acute{n}_r = 6$, dan

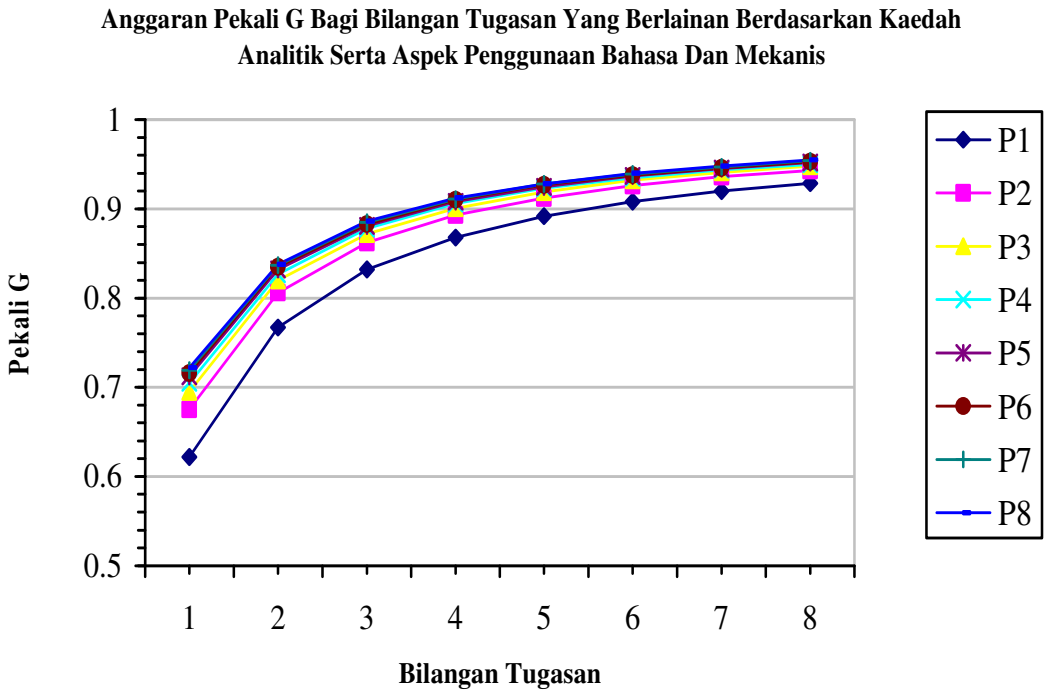
$\acute{n}_t$ = 3 berubah menjadi $\acute{n}_t$ = 4, peratusan pekali G untuk bilangan tugasan meningkat sebanyak 3.1% (.910 – .883) manakala apabila $\acute{n}_t$ = 6, dan $\acute{n}_r$ = 3 berubah kepada $\acute{n}_r$ = 4, pekali G untuk bilangan pemeriksa hanya meningkat sebanyak 0.3% (.935 – .932).

Jadual 4.14

Keputusan Pekali G Bagi Bilangan Tugasan Dan Bilangan Pemeriksa Yang Berlainan Berdasarkan Kaedah Holistik Serta Aspek Penggunaan Bahasa Dan Mekanis

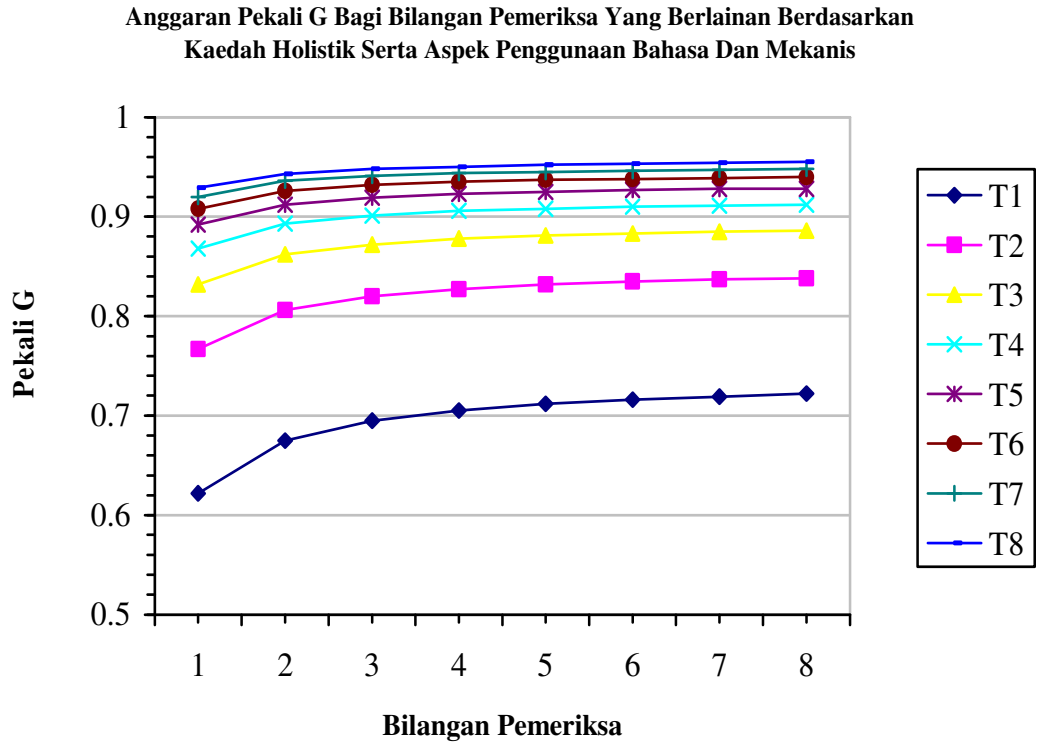
Bilangan Pemeriksa ($\acute{n}_r$)	Bilangan Tugasan ($\acute{n}_t$)							
	1	2	3	4	5	6	7	8
1	.622	.767	.832	.868	.892	.908	.920	.929
2	.675	.806	.862	.893	.912	.926	.936	.943
3	.695	.820	.872	.901	.919	.932	.941	.948
4	.705	.827	.878	.906	.923	.935	.944	.950
5	.712	.832	.881	.908	.925	.937	.945	.952
6	.716	.835	.883	.910	.927	.938	.946	.953
7	.719	.837	.885	.911	.928	.939	.947	.954
8	.722	.838	.886	.912	.928	.940	.948	.955

Nota. Berdasarkan data asal (n_p = *120,* n_t = *2,* n_r = *3*) dengan reka bentuk separa tersarang $p \times (r{:}t)$.



Rajah 4.7. Pekali G bagi bilangan tugasan yang berlainan berdasarkan kaedah holistik serta aspek penggunaan bahasa dan mekanis.

Nota. P1= pemeriksa pertama; P2 = pemeriksa kedua dan seterusnya.



Rajah 4.8. Pekali G bagi bilangan pemeriksa yang berlainan berdasarkan kaedah holistik serta aspek penggunaan bahasa dan mekanis.

Nota. T1= tugasan pertama; T2 = tugasan kedua dan seterusnya.

Untuk meninjau perubahan pekali G berdasarkan kombinasi bilangan tugasan dan pemeriksa bagi prosedur ini, pertambahan bilangan tugasan nampaknya mempunyai impak yang lebih besar daripada bilangan pemeriksa terhadap kebergantungan skor. Misalnya, bagi senario tugasan dan pemeriksa tunggal untuk prosedur pemarkahan ini, pekali G ialah .622. Apabila kombinasi $\acute{n}_t = 1$, $\acute{n}_r = 2$ digunakan, pekali G meningkat 8.5% (.675 –.622). Namun, apabila berdasarkan kombinasi $\acute{n}_t = 2$, $\acute{n}_r = 1$, pekali G meningkat lebih tinggi lagi iaitu .767 atau 23.3%. Pola yang sama boleh dilihat menerusi kombinasi $\acute{n}_t = 3$, $\acute{n}_r = 4$ (.878) berbanding dengan kombinasi $\acute{n}_t = 4$, $\acute{n}_r = 3$ (.901); kombinasi $\acute{n}_t = 4$, $\acute{n}_r = 5$ (.908) berbanding dengan kombinasi $\acute{n}_t = 5$, $\acute{n}_r = 4$ (.923) dan seterusnya. Malah dengan satu pemeriksa, enam tugasan diperlukan untuk mencecah kriteria .90 yang dihajati. Manakala dengan satu tugasan, walaupun lapan pemeriksa digunakan, pekali G hanya mencecah .72. Sementara itu, pekali G untuk reka bentuk kajian asal bagi prosedur ini iaitu $\acute{n}_t = 2$, $\acute{n}_r = 3$ adalah .82. Sekiranya kombinasi $\acute{n}_t = 3$, $\acute{n}_r = 2$ digunakan, kepersisan pengukuran boleh diperbaiki lagi kerana pekali G akan meningkat kepada .86. Kombinasi optimum bagi bilangan tugasan dan bilangan pemeriksa berdasarkan kriteria .90 bagi prosedur pemarkahan ini adalah $\acute{n}_t = 4$, $\acute{n}_r = 3$. Kombinasi lain yang boleh dipertimbangkan adalah $\acute{n}_t = 5$, $\acute{n}_r = 2$ atau $\acute{n}_t = 6$, $\acute{n}_r = 1$.

Jadual 4.15 serta Rajah 4.9 dan Rajah 4.10 memaparkan anggaran pekali G ($E\rho^2$) melalui operasi pelbagai kajian D untuk reka bentuk separa tersarang $p \times (r{:}t)$ bagi bilangan tugasan dan bilangan pemeriksa yang berlainan berdasarkan kaedah analitik serta aspek penggunaan bahasa dan mekanis. Dalam prosedur pemarkahan ini, pertambahan bilangan tugasan secara relatif mempunyai impak yang lebih besar terhadap kebergantungan skor. Namun impak untuk pertambahan bilangan pemeriksa

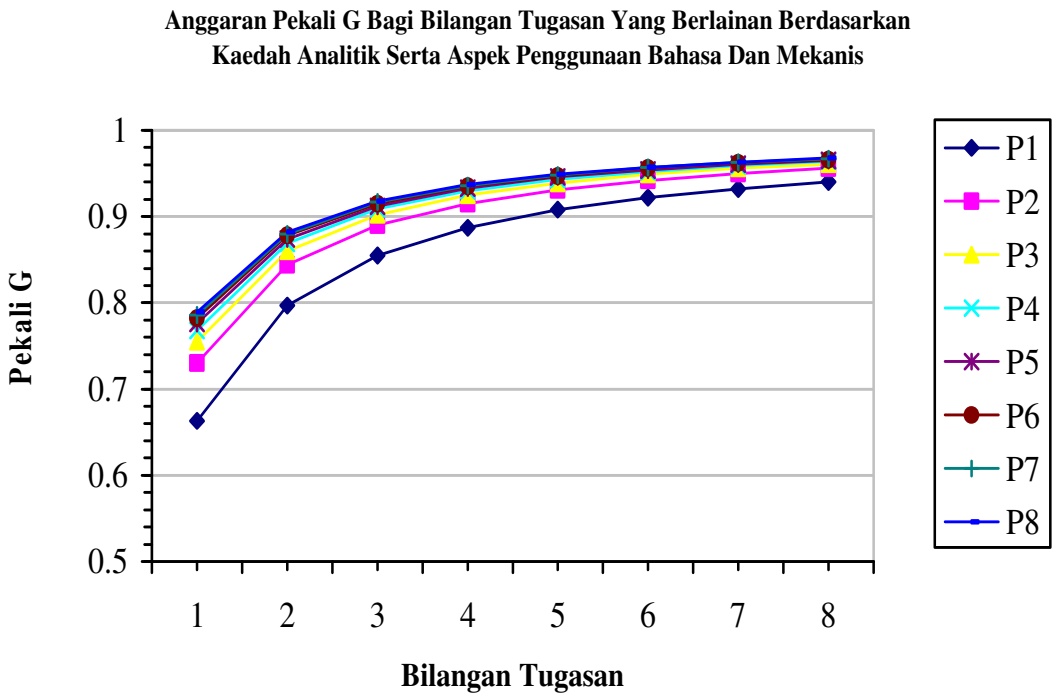
terhadap kebergantungan skor secara relatif adalah kecil. Misalnya, berdasarkan senario pemeriksa tunggal, pekali G meningkat secara drastik iaitu daripada .663 kepada .797 atau sebanyak 20.2% apabila $\acute{n}_t = 1$ berubah kepada $\acute{n}_t = 2$. Manakala berdasarkan senario tugasan tunggal, pekali G hanya meningkat sebanyak .067 (.730–.663) atau 10.1% apabila $\acute{n}_t = 1$ berubah kepada $\acute{n}_t = 2$. Manakala bagi pertambahan bilangan tugasan yang selanjutnya daripada tiga hingga lapan iaitu pertambahan sebanyak enam tugasan, memperlihatkan pekali G meningkatkan sebanyak .085 (.940 –.855) atau 9.9%. Namun, untuk pertambahan bilangan pemeriksa dengan kadar yang sama, pekali G hanya meningkat sebanyak .034 (.789 – .755) atau 4.5%. Ini juga menunjukkan bahawa terdapat pulangan menurun dalam pekali G apabila lebih banyak tugasan atau pemeriksa digunakan. Pola yang sama boleh didapati apabila bilangan tugasan atau bilangan pemeriksa yang seterusnya digunakan, pekali G akan meningkat dengan cepat pada peringkat awalnya terutamanya bagi bilangan tugasan dan bilangan pemeriksa yang kedua

Jadual 4.15

Keputusan Pekali G Bagi Bilangan Tugasan Dan Bilangan Pemeriksa Yang Berlainan Berdasarkan Kaedah Analitik Serta Aspek Penggunaan Bahasa Dan Mekanis

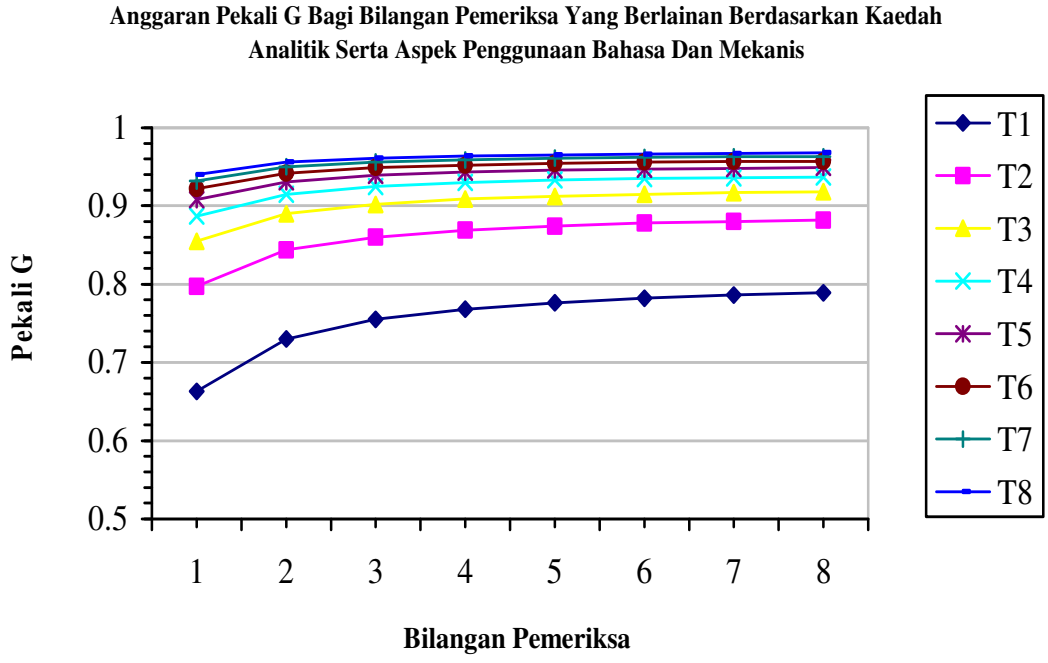
Bilangan Pemeriksa ($\acute{n}_r$)	Bilangan Tugasan ($\acute{n}_t$)							
	1	2	3	4	5	6	7	8
1	.663	.797	.855	.887	.908	.922	.932	.940
2	.730	.844	.890	.915	.931	.942	.950	.956
3	.755	.860	.902	.925	.939	.949	.956	.961
4	.768	.869	.909	.930	.943	.952	.959	.964
5	.776	.874	.912	.933	.946	.954	.961	.965
6	.782	.878	.915	.935	.947	.956	.962	.966
7	.786	.880	.917	.936	.948	.957	.963	.967
8	.789	.882	.918	.937	.949	.957	.963	.968

Nota. Berdasarkan data asal ($n_p = 120$, $n_t = 2$, $n_r = 3$)dengan reka bentuk separa tersarang $p \times (r{:}t)$.



Rajah 4.9. Pekali G bagi bilangan tugasan yang berlainan berdasarkan kaedah analitik serta aspek penggunaan bahasa dan mekanis.

Nota. P1= pemeriksa pertama; P2 = pemeriksa kedua dan seterusnya.



Rajah 4.10. Pekali G bagi bilangan pemeriksa yang berlainan berdasarkan kaedah analitik serta aspek penggunaan bahasa dan mekanis.

Nota. T1= tugasan pertama; T2 = tugasan kedua dan seterusnya.

dan pertambahan pekali G akan berkurangan selepas itu (Rajah 4.9 dan Rajah 4.10). Namun begitu, peratusan peningkatan pekali G bagi bilangan tugasan adalah lebih tinggi berbanding dengan bilangan pemeriksa untuk semua keadaan pertambahan bilangan dengan kadar yang sama.

Untuk meninjau perubahan pekali G berdasarkan kombinasi bilangan tugasan dan pemeriksa bagi prosedur ini, seperti juga prosedur-prosedur pemarkahan yang lain, pertambahan bilangan tugasan nampaknya mempunyai impak yang lebih besar daripada pertambahan bilangan pemeriksa terhadap kebergantungan skor. Misalnya, bagi senario tugasan dan pemeriksa tunggal untuk prosedur pemarkahan ini, pekali G ialah .66. Apabila kombinasi $\acute{n}_t = 1$, $\acute{n}_r = 2$ digunakan, pekali G meningkat kepada .73. Walau bagaimanapun, apabila kombinasi $\acute{n}_t = 2$, $\acute{n}_r = 1$, pekali G meningkat lebih tinggi lagi iaitu .80. Pola yang sama boleh dilihat menerusi kombinasi $\acute{n}_t = 3$, $\acute{n}_r = 4$ (.909) berbanding dengan kombinasi $\acute{n}_t = 4$, $\acute{n}_r = 3$ (.925); kombinasi $\acute{n}_t = 4$, $\acute{n}_r = 5$ (.933) berbanding dengan kombinasi $\acute{n}_t = 5$, $\acute{n}_r = 4$ (.943) dan seterusnya. Malah dengan satu pemeriksa, lima tugasan diperlukan untuk mencecah kriteria .90 yang dihajati. Manakala dengan satu tugasan, walaupun lapan pemeriksa digunakan, pekali G hanya mencecah .79. Sementara itu, pekali G untuk reka bentuk kajian asal bagi prosedur ini iaitu $\acute{n}_t = 2$, $\acute{n}_r = 3$ ialah .86, sekiranya kombinasi $\acute{n}_t = 3$, $\acute{n}_r = 2$ digunakan, kepersisan pengukuran boleh diperbaiki lagi kerana pekali G akan meningkat kepada .89. Dalam prosedur pemarkahan ini, kriteria .90 boleh dicapai menerusi kombinasi optimum tiga buah tugasan dan tiga orang pemeriksa ($\acute{n}_t = 3$, $\acute{n}_r = 3$). Pilihan pemadanan bilangan tugasan dan bilangan pemeriksa yang lain adalah $\acute{n}_t = 4$, $\acute{n}_r = 2$ atau $\acute{n}_t = 5$, $\acute{n}_r = 1$.

Secara kesimpulan, rumusan dapatan kajian adalah seperti berikut:

Perosedur Pemarkahan	Impak Terhadap kebergantungan Skor		
	Bilangan Tugasan	Bilangan Pemeriksa	Gabungan Bilangan Tugasan Dan Pemeriksa
Kaedah holistik serta aspek kandungan dan organisasi	• Peningkatan drastik pekali G kira-kira .159 (.710–.551) atau 28.9% apabila $\acute{n}_t$= 1 berubah kpd $\acute{n}_t$=2 berdasarkan $\acute{n}_r$= 1. Pertambahan $\acute{n}_t$ yang seterusnya memaparkan % peningkatan pekali yang semakin menurun. • Pola yang sama apabila $\acute{n}_r$ bertambah.	• Peningkatan pekali G yang lebih besar iaitu kira-kira .055 (.606 – .551) atau 10.0% ketika $\acute{n}_r$= 1 berubah kpd $\acute{n}_r$= 2 berdasarkan $\acute{n}_t$= 1. Pertambahan $\acute{n}_r$ yang seterusnya memaparkan % peningkatan pekali yang semakin menurun. • Pola yang sama apabila $\acute{n}_t$ bertambah.	• Impak pertambahan bilangan tugasan adalah lebih besar daripada pertambahan bilangan pemeriksa. • Pilihan pemadanan optimum pada kriteria .90 adalah $\acute{n}_t$ = 5, $\acute{n}_r$ = 4, $\acute{n}_t$ = 6, $\acute{n}_r$ = 2 atau $\acute{n}_t$ = 7, $\acute{n}_r$ = 1.
Kaedah analitik serta aspek kandungan dan organisasi	• Peningkatan drastik pekali G kira-kira .15 (.75 – .60) atau 25.0% apabila $\acute{n}_t$ = 1 berubah kpd $\acute{n}_t$ =2 berdasarkan $\acute{n}_r$ = 1. Pertambahan $\acute{n}_t$ yang seterusnya memaparkan % peningkatan pekali yang semakin menurun. • Pola yang sama apabila $\acute{n}_r$ bertambah.	• Peningkatan pekali G yang lebih besar iaitu kira-kira .053 (.653 – .600) atau 8.8% apabila $\acute{n}_r$ = 1 berubah kepada $\acute{n}_r$ = 2 berdasarkan $\acute{n}_t$ = 1. Pertambahan $\acute{n}_r$ yang seterusnya memaparkan % peningkatan pekali yang semakin menurun. • Pola yang sama apabila $\acute{n}_t$ bertambah.	• Impak pertambahan bilangan tugasan adalah lebih besar daripada pertambahan bilangan pemeriksa. • Pilihan pemadanan optimum pada kriteria .90 adalah $\acute{n}_t$ = 4, $\acute{n}_r$ = 4, $\acute{n}_t$ = 5, $\acute{n}_r$ = 2 atau $\acute{n}_t$ = 6, $\acute{n}_r$ = 1.
Kaedah holistik serta aspek penggunaan bahasa dan mekanis	• Peningkatan drastik pekali G kira-kira .145 (.767–.622) atau 23.3% apabila $\acute{n}_t$ = 1 berubah kpd $\acute{n}_t$ =2 berdasarkan $\acute{n}_r$ = 1. Pertambahan $\acute{n}_t$ yang seterusnya memaparkan % peningkatan pekali yang semakin menurun. • Pola yang sama apabila $\acute{n}_r$ bertambah.	• Peningkatan pekali G yang lebih besar iaitu kira-kira .053 (.675 – .622) atau 8.5% ketika $\acute{n}_r$= 1 berubah kpd $\acute{n}_r$= 2 berdasarkan $\acute{n}_t$= 1. Pertambahan $\acute{n}_r$ yang seterusnya memaparkan % peningkatan pekali yang semakin menurun. • Pola yang sama apabila $\acute{n}_t$ bertambah.	• Impak pertambahan bilangan tugasan adalah lebih besar daripada pertambahan bilangan pemeriksa. • Pilihan pemadanan optimum pada kriteria .90 adalah $\acute{n}_t$ = 4, $\acute{n}_r$ = 3, $\acute{n}_t$ = 5, $\acute{n}_r$ = 2 atau $\acute{n}_t$ = 6, $\acute{n}_r$ = 1.
Kaedah analitik serta aspek penggunaan bahasa dan mekanis	• Peningkatan drastik pekali G kira-kira .134 (.797– .663) atau 20.2% apabila $\acute{n}_t$ = 1 berubah kpd $\acute{n}_t$ =2 berdasarkan $\acute{n}_r$ = 1. Pertambahan $\acute{n}_t$ seterusnya memaparkan % peningkatan pekali yang semakin menurun. • Pola yang sama apabila $\acute{n}_r$ bertambah.	• Peningkatan pekali G yang lebih besar iaitu kira-kira .067 (.730 – .663) atau 10.1% ketika $\acute{n}_r$= 1 berubah kpd $\acute{n}_r$ = 2 berdasarkan $\acute{n}_t$= 1. Pertambahan $\acute{n}_r$ seterusnya memaparkan % peningkatan pekali yang semakin menurun. • Pola yang sama apabila $\acute{n}_t$ bertambah.	• Impak pertambahan bilangan tugasan adalah lebih besar daripada pertambahan bilangan pemeriksa. • Pilihan pemadanan optimum pada kriteria .90 adalah $\acute{n}_t$ = 3, $\acute{n}_r$ = 3, $\acute{n}_t$ = 4, $\acute{n}_r$ = 2 atau $\acute{n}_t$ = 5, $\acute{n}_r$ = 1.

Berdasarkan analisis hasil keputusan di atas, maka terjawablah soalan kajian 3(a) iaitu berdasarkan kerangka kajian D, apakah impak bagi bilangan tugasan karangan, bilangan pemeriksa dan gabungan kedua-duanya terhadap kebergantungan skor bagi prosedur pemarkahan iaitu:

(v) kaedah holistik serta aspek kandungan dan organisasi

(vi) kaedah analitik serta aspek kandungan dan organisasi

(vii) kaedah holistik serta aspek penggunaan bahasa dan mekanis

(viii) kaedah analitik serta aspek penggunaan bahasa dan mekanis

Analisis dapatan kajian seterusnya adalah untuk menjawab soalan kajian 3(b) iaitu sejauh manakah impak gabungan bilangan tugasan karangan dan bilangan pemeriksa bagi prosedur pemarkahan yang berlainan terhadap kebergantungan skor berdasarkan kerangka kajian D.

Jadual 4.16 melaporkan hasil keputusan perubahan pekali G berdasarkan reka bentuk separa tersarang $p \times (R: T)$ kajian D model rawak untuk kombinasi bilangan tugasan dan bilangan pemeriksa satu hingga lapan bagi prosedur pemarkahan yang berlainan. Keputusan bagi operasi pelbagai kajian D berdasarkan reka bentuk tersebut boleh dilihat dalam Lampiran T 1 hingga Lampiran T 4 dan telah dirumuskan dalam Jadual 4.16.

Dapatan kajian dalam Jadual 4.16 menunjukkan bahawa berdasarkan gabungan bilangan tugasan dan pemeriksa tunggal, prosedur pemarkahan yang berlainan menunjukkan nilai pekali G yang berbeza-beza iaitu antara .55 hingga .66.

Ini bermakna korelasi antara skor yang diperoleh dengan kebolehan sebenar murid adalah kira-kira antara .74 hingga .81 (antara √.55 hingga √.66) berdasarkan prosedur pemarkahan yang berlainan. Antaranya, berdasarkan prosedur pemarkahan untuk kaedah analitik serta aspek penggunaan bahasa dan mekanis, nilai pekali G paling tinggi (.66) manakala dalam kaedah holistik serta aspek kandungan dan organisasi pula nilainya adalah paling rendah (.55). Ini menunjukkan berdasarkan pemadanan bilangan tugasan dan bilangan pemeriksa tunggal, tidak ada satu prosedur pemarkahan memaparkan kebergantungan skor yang menyakinkan dalam mentaksir kemahiran menulis karangan murid sekiranya pekali G ditetapkan pada kriteria .90.

Untuk senario tugasan tunggal, walaupun lapan orang pemeriksa digunakan, masih tidak ada satu prosedur pemarkahan yang dapat mencapai nilai pekali pada tahap .80. Sebaliknya, berdasarkan senario pemeriksa tunggal dan lapan buah tugasan, semua prosedur pemarkahan dapat memenuhi kriteria .90. Secara terperinci, untuk mencecah pekali pada .90 berdasarkan senario pemeriksa tunggal, kaedah analitik serta aspek penggunaan bahasa dan mekanis memerlukan lima buah tugasan ($\acute{n}_t = 5$, $\acute{n}_r = 1$) manakala kaedah holistik serta aspek kandungan dan organisasi pula memerlukan tujuh buah tugasan ($\acute{n}_t = 7$, $\acute{n}_r = 1$). Sementara itu, kaedah holistik serta aspek penggunaan bahasa dan mekanis, dan kaedah analitik serta aspek kandungan dan organisasi memerlukan enam buah tugasan masing-masing ($\acute{n}_t = 6$, $\acute{n}_r = 1$). Ini menunjukkan pertambahan bilangan tugasan adalah lebih berkesan berbanding dengan pertambahan pemeriksa untuk mencapai pekali G yang tinggi.

Jadual 4.16

Analisis Keadaan Perubahan Pekali G Bagi Bilangan Tugasan Dan Bilangan Pemeriksa Yang Berlainan Berdasarkan Prosedur Pemarkahan Yang Berbeza

Bilangan Tugasan	Bilangan Pemeriksa	Aspek Kandungan Dan Organisasi		Aspek Penggunaan Bahasa Dan Mekanis	
		Kaedah Holistik	Kaedah Analitik	Kaedah Holistik	Kaedah Analitik
1	1	.551	.600	.622	.663
	2	.606	.653	.675	.730
	3	.627	.673	.695	.755
	4	.638	.683	.705	.768
	5	.645	.689	.712	.776
	6	.649	.694	.716	.782
	7	.652	.697	.719	.786
	8	.655	.699	.722	.789
2	1	.710	.750	.767	.797
	2	.755	.790	.806	.844
	3	.771	.804	.820	.860
	4	.779	.812	.827	.869
	5	.784	.816	.832	.874
	6	.787	.819	.835	.878
	7	.790	.821	.837	.880
	8	.792	.823	.838	.882
3	1	.786	.818	.832	.855
	2	.822	.849	.862	.890
	3	.834	.860	.872	.902
	4	.841	.866	.878	.909
	5	.845	.869	.881	.912
	6	.847	.872	.883	.915
	7	.849	.873	.885	.917
	8	.851	.875	.886	.918
4	1	.831	.857	.868	.887
	2	.860	.883	.893	.915
	3	.870	.892	.901	.925
	4	.876	.896	.906	.930
	5	.879	.899	.908	.933
	6	.881	.901	.910	.935
	7	.882	.902	.911	.936
	8	.884	.903	.912	.937
5	1	.860	.882	.892	.908
	2	.885	.904	.912	.931
	3	.894	.911	.919	.939
	4	.898	.915	.923	.943
	5	.901	.917	.925	.946
	6	.902	.919	.927	.947
	7	.904	.920	.928	.948
	8	.905	.921	.928	.949
6	1	.880	.900	.908	.922
	2	.902	.919	.926	.942
	3	.910	.925	.932	.949
	4	.914	.928	.935	.952
	5	.916	.930	.937	.954
	6	.917	.931	.938	.956
	7	.918	.932	.939	.957
	8	.919	.933	.940	.957
7	1	.896	.913	.920	.932
	2	.915	.929	.936	.950
	3	.922	.935	.941	.956
	4	.925	.938	.944	.959
	5	.927	.940	.945	.961
	6	.928	.941	.946	.962
	7	.929	.941	.947	.963
	8	.930	.942	.948	.963
8	1	.908	.923	.929	.940
	2	.925	.938	.943	.956
	3	.931	.943	.948	.961
	4	.934	.945	.950	.964
	5	.936	.947	.952	.965
	6	.937	.948	.953	.966
	7	.938	.948	.954	.967
	8	.938	.949	.955	.968

Seperti yang ditunjukkan dalam analisis hasil kajian terlebih dahulu, peningkatan peratusan pekali G yang lebih besar akan diperoleh apabila bilangan tugasan atau bilangan pemeriksa bertambah daripada satu kepada dua tanpa mengira prosedur pemarkahan yang digunakan. Ini terutamanya bagi pertambahan bilangan tugasan daripada satu kepada dua. Secara analisis, pertambahan bilangan tugasan daripada satu kepada dua berdasarkan pemeriksa tunggal akan meningkatkan peratusan pekali G antara 20.2% hingga 28.9% manakala pertambahan bilangan pemeriksa daripada satu kepada dua berdasarkan tugasan tunggal pula akan meningkatkan peratusan pekali G antara 8.5% hingga 10.1%, merentas semua prosedur pemarkahan (Jadual 4.16). Antaranya, kaedah holistik serta aspek kandungan dan organisasi mengalami peningkatan peratusan pekali G yang paling besar bagi pertambahan bilangan tugasan daripada satu kepada dua (28.9%) manakala kaedah analitik serta aspek penggunaan bahasa dan mekanis pula memaparkan peningkatan peratusan pekali G yang paling besar bagi pertambahan bilangan pemeriksa daripada satu kepada dua (10.1%). Ini secara langsung juga menunjukkan bahawa pertambahan dalam bilangan tugasan mempunyai kesan yang lebih ketara berbanding dengan pertambahan dalam bilangan pemeriksa bagi meningkatkan kebergantungan skor merentas semua prosedur pemarkahan.

Namun begitu, gabungan bilangan tugasan dan bilangan pemeriksa akan mengalami pulangan menurun dari segi pekali G apabila bilangan semakin bertambah. Ini boleh dilihat menerusi perubahan peratusan pekali G berdasarkan gabungan bilangan tugasan dan pemeriksa (Jadual 4.17). Misalnya bagi kaedah holistik serta aspek kandungan dan organisasi, pertambahan bilangan tugasan daripada satu kepada dua dengan pemeriksa tunggal ($\acute{n}_t = 2$, $\acute{n}_r = 1$) memperlihatkan

pekali G meningkat kira-kira 28.9%, dan bagi gabungan $\acute{n}_t = 3$, $\acute{n}_r = 1$ pula, pekali G meningkat kira-kira 42.7%. Manakala apabila gabungan $\acute{n}_t = 8$, $\acute{n}_r = 1$ digunakan, pekali G meningkat kira-kira 64.8%. Ini bermakna pertambahan lima buah tugasan yang seterusnya hanya melibatkan peningkatan pekali G kira-kira 22.1% (64.8% – 42.7%). Sedangkan berdasarkan gabungan $\acute{n}_t = 2$, $\acute{n}_r = 1$, pekali G sudah meningkat kira-kira 28.9%. Secara ringkas, berdasarkan peningkatan peratusan pekali G bagi gabungan $\acute{n}_t = 2$, $\acute{n}_r = 1$ berbanding dengan pertambahan lima buah tugasan yang seterusnya (gabungan $\acute{n}_t = 3$, $\acute{n}_r = 1$ hingga gabungan $\acute{n}_t = 8$, $\acute{n}_r = 1$), kaedah analitik serta aspek kandungan dan organisasi memaparkan peningkatan pekali G kira-kira 25.0% berbanding dengan kira-kira 17.5% (53.8% – 36.3%); kaedah holistik serta aspek penggunaan bahasa dan mekanis adalah kira-kira 23.3% berbanding dengan kira-kira 15.6% (49.4% – 33.8%) manakala kaedah analitik serta aspek penggunaan bahasa dan mekanis adalah kira-kira 20.2% berbanding dengan kira-kira 12.8% (41.8% – 29.0%). Ini menunjukkan bahawa pekali G telah mengalami pulangan menurun apabila bilangan tugasan dalam sesuatu gabungan semakin bertambah.

Keadaan yang sama juga berkaku pada pertambahan dari segi bilangan pemeriksa. Bagi kaedah holistik serta aspek kandungan dan organisasi, berdasarkan gabungan $\acute{n}_t = 1$ $\acute{n}_r = 2$, pekali meningkat kira-kira 10.0% manakala bagi gabungan $\acute{n}_t = 1$ $\acute{n}_r = 8$ iaitu pertambahan seramai enam pemeriksa, pekali hanya meningkat kira-kira 8.9% (18.9% – 10.0%). Begitu juga bagi kaedah analitik serta aspek kandungan dan organisasi, pertambahan enam pemeriksa yang selanjutnya hanya meningkatkan pekali G kira-kira 7.7% (16.5% – 8.8%) berbanding dengan 8.8% bagi gabungan $\acute{n}_t = 1$ $\acute{n}_r = 2$; kaedah holistik serta aspek penggunaan bahasa dan mekanis iaitu kira-kira 7.6% (16.1% – 8.5%) berbanding dengan 8.5%; dan kaedah analitik serta aspek

penggunaan bahasa dan mekanis iaitu kira-kira 8.9% (19.0% – 10.1%) berbanding dengan 10.1% . Keadaan ini juga menunjukkan pertambahan bilangan pemeriksa adalah kurang berkesan berbanding dengan pertambahan bilangan tugasan dalam meningkatkan kebergantungan skor.

Jadual 4.17

Keadaan Perubahan Peratusan Pekali G Berdasarkan Perubahan Gabungan Bilangan Tugasan dan Bilangan Pemeriksa Bagi Prosedur Pemarkahan Yang Berlainan

Gabungan Bil Tugasan Dan Bil Pemeriksa	Perubahan Peratusan Pekali G (%)			
	Kaedah Holistik Serta Aspek Kandungan Dan Organisasi	Kaedah Analitik Serta Aspek Kandungan Dan Organisasi	Kaedah Holistik Serta Aspek Penggunaan Bhs Dan Mekanis	Kaedah Analitik Serta Aspek Penggunaan Bhs Dan Mekanis
$\acute{n}_t = 1, \acute{n}_r = 2$	10.0	8.8	8.5	10.1
$\acute{n}_t = 1, \acute{n}_r = 8$	18.9	16.5	16.1	19.0
$\acute{n}_t = 2, \acute{n}_r = 1$	28.9	25.0	23.3	20.2
$\acute{n}_t = 2, \acute{n}_r = 2$	37.0	31.7	29.6	27.3
$\acute{n}_t = 3, \acute{n}_r = 1$	42.7	36.3	33.8	29.0
$\acute{n}_t = 3, \acute{n}_r = 3$	51.4	43.3	40.2	36.1
$\acute{n}_t = 4, \acute{n}_r = 4$	59.0	49.3	45.7	40.3
$\acute{n}_t = 8, \acute{n}_r = 1$	64.8	53.8	49.4	41.8
$\acute{n}_t = 8, \acute{n}_r = 8$	70.2	58.2	53.5	46.0

Nota. Anggaran perubahan peratusan pekali adalah berasaskan gabungan tugasan dan pemeriksa tunggal ($\acute{n}_t = 1, \acute{n}_r = 1$)

Begitu juga bagi gabungan bilangan tugasan dan pemeriksa dengan kadar yang sama, pertambahan bilangan kedua-duanya akan menyebabkan pulangan menurun kepada peratusan pekali G. Misalnya, bagi kaedah holistik serta aspek kandungan dan organisasi, apabila gabungan tugasan dan pemeriksa tunggal ($\acute{n}_t = 1$, $\acute{n}_r = 1$) berubah kepada $\acute{n}_t = 2$, $\acute{n}_r = 2$, pekali G meningkat kira-kira 37.0%; apabila $\acute{n}_t$

= 2, $\acute{n}_r$ = 2 berubah kepada gabungan $\acute{n}_t$ = 3, $\acute{n}_r$ = 3, pekali G meningkat kira-kira 14.4% (51.4 – 37.0%) manakala apabila $\acute{n}_t$ = 4, $\acute{n}_r$ = 4 berubah kepada gabungan $\acute{n}_t$ = 8, $\acute{n}_r$ = 8, pekali G hanya meningkat kira-kira 11.2% (70.2 – 59.0%). Keadaan yang sama juga berlaku pada prosedur pemarkahan yang lain (Jadual 4.17). Ini menunjukkan gabungan bilangan tugasan dan bilangan pemeriksa yang semakin meningkat akan menyebabkan peningkatan pekali G yang semakin berkurangan.

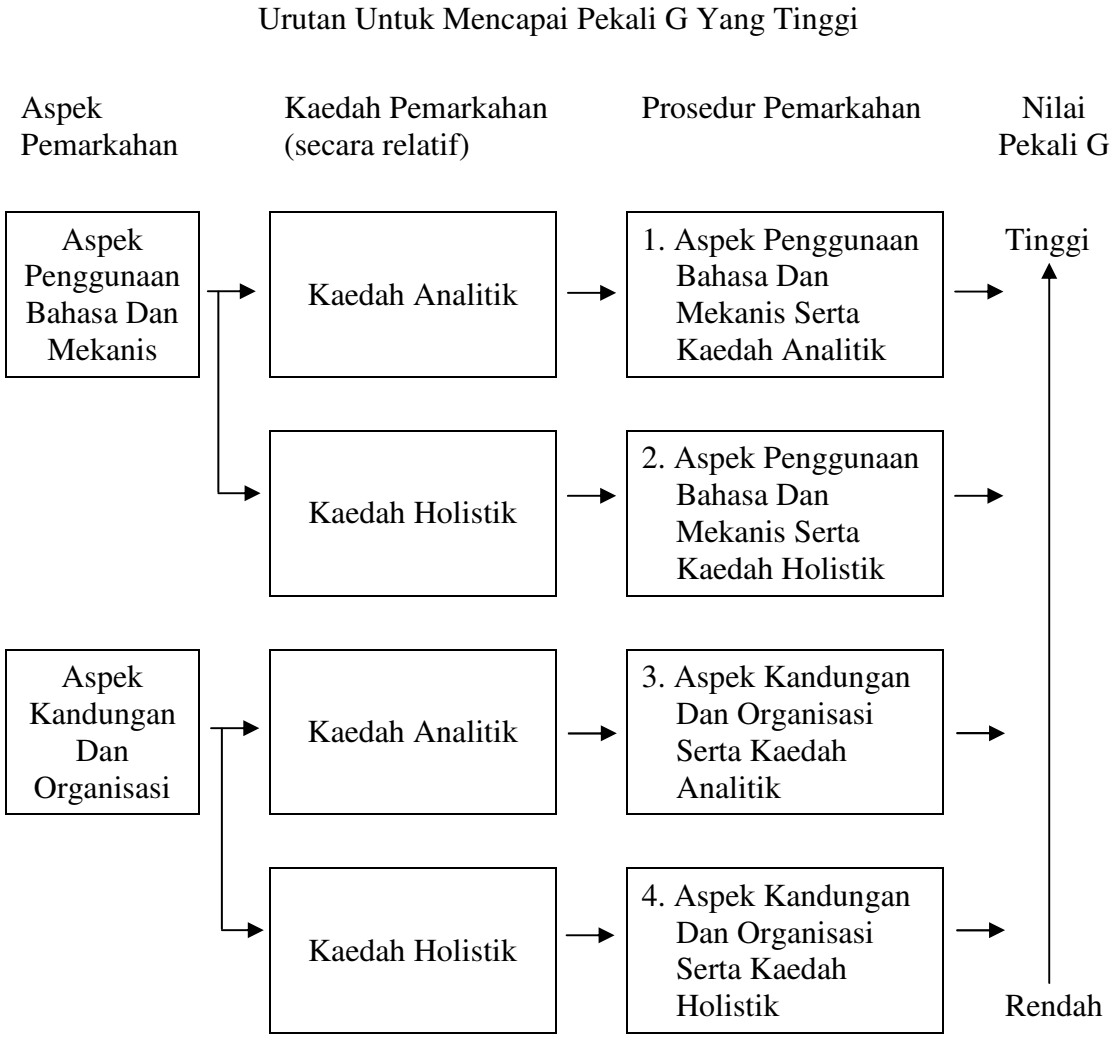
Dapatan kajian menunjukkan pada keseluruhannya pertambahan bilangan pemeriksa yang lebih daripada empat orang tidak banyak menyumbang terhadap kebergantungan skor manakala pertambahan tugasan bilangan yang sama masih mempunyai ruang untuk meningkatkan nilai pekali G merentas semua prosedur pemarkahan (Jadual 4.16). Misalnya untuk kaedah analitik serta aspek penggunaan bahasa dan mekanis, apabila bilangan pemeriksa bertambah daripada empat kepada lima orang dengan empat buah tugasan digunakan ($\acute{n}_t$ = 4, $\acute{n}_r$ = 4; $\acute{n}_t$ = 4, $\acute{n}_r$ = 5), pekali G hanya berubah daripada .930 kepada .933. Manakala apabila bilangan tugasan ditambah kepada lima ($\acute{n}_t$ = 4, $\acute{n}_r$ = 4; $\acute{n}_t$ = 5, $\acute{n}_r$ = 4), pekali G meningkat kepada .943.

Dapatan kajian dengan jelas menunjukkan bahawa aspek penggunaan bahasa dan mekanis lebih mudah mencecah nilai pekali yang tinggi berbanding dengan aspek kandungan dan organisasi berdasarkan gabungan bilangan tugasan dan bilangan pemeriksa yang sama tanpa mengira sama ada kaedah holistik atau kaedah analitik digunakan (Jadual 4.16). Ini dapat dilihat melalui senario gabungan tugasan dan pemeriksa tunggal ($\acute{n}_t$ = 1, $\acute{n}_r$ = 1), pekali G bagi aspek penggunaan bahasa dan mekanis adalah lebih tinggi iaitu .62 (kaedah holistik) dan .66 (kaedah analitik)

masing-masing berbanding dengan aspek kandungan dan organisasi iaitu .55 (kaedah holistik) dan .60 (kaedah analitik). Senario ini menjadi lebih ketara lagi bagi aspek penggunaan bahasa dan mekanis yang digabungkan dengan kaedah analitik. Prosedur pemarkahan ini merupakan prosedur yang paling mudah mencapai kriteria .90 iaitu apabila sampel karangan meningkat menjadi tiga dan tiga orang pemeriksa digunakan ($\acute{n}_t = 3$, $\acute{n}_r = 3$). Secara teliti, pekali G bagi prosedur pemarkahan ini akan bertambah dengan bertambahnya bilangan tugasan dan bilangan pemeriksa berdasarkan lingkungan bilangan tugasan dan bilangan pemeriksa satu hingga lapan. Keadaan nilai pekalinya akan sentiasa lebih tinggi jika dibandingkan dengan prosedur pemarkahan yang lain. Walaupun kaedah holistik dengan aspek penggunaan bahasa dan mekanis juga dapat mencecah pekali G pada kriteria .90 apabila bilangan pemeriksa bertambah kepada tiga, namun begitu ia memerlukan empat tugasan untuk mencapai kriteria tersebut ($\acute{n}_t = 4$, $\acute{n}_r = 3$). Justeru itu, keberkesanan dan tahap kepersisan pengukuran bagi prosedur tersebut masih berada di bawah kedudukan kaedah analitik yang berdasarkan aspek pemarkahan yang sama. Manakala aspek kandungan dan organisasi serta aspek analitik memerlukan pemadanan bilangan tugasan dan bilangan pemeriksa yang sama iaitu empat ($\acute{n}_t = 4$, $\acute{n}_r = 4$) untuk mencecah pekali G pada tahap .90. Sementara itu, berdasarkan aspek pemarkahan yang sama, kaedah holistik memerlukan pemadanan lima buah tugasan karangan dan empat orang pemeriksa ($\acute{n}_t = 5$, $\acute{n}_r = 4$) untuk mencapai kriteria .90.

Secara ringkas, aspek penggunaan bahasa dan mekanis memerlukan pemadanan bilangan tugasan dan bilangan pemeriksa yang paling kurang untuk mencapai nilai pekali G yang tinggi berbanding dengan aspek kandungan dan organisasi. Dari segi kaedah pemarkahan pula, kaedah analitik secara relatif

memerlukan pemadanan bilangan tugasan dan bilangan pemeriksa yang kurang untuk mencapai nilai pekali G yang lebih tinggi berbanding dengan kaedah holistik. Dari perspektif prosedur pemarkahan, kaedah analitik serta aspek penggunaan bahasa dan mekanis merupakan prosedur pemarkahan yang memerlukan pemadanan bilangan tugasan dan bilangan pemeriksa yang paling sedikit untuk mencecah pekali yang tinggi dalam kajian ini. Sementara itu, kebolehpercayaan skor bagi prosedur pemarkahan yang menggunakan kaedah holistik serta aspek penggunaan bahasa dan mekanis adalah sedikit lebih baik daripada kaedah analitik serta aspek kandungan dan organisasi berdasarkan gabungan bilangan tugasan dan pemeriksa yang sama.



Rajah 4.11. Urutan untuk mencapai nilai pekali G yang tinggi berdasarkan gabungan bilangan tugasan dan pemeriksa yang paling sedikit

Manakala kaedah holistik serta aspek kandungan dan organisasi merupakan prosedur pemarkahan yang paling kurang memuaskan dari segi pencapaian pekali G yang tinggi. Secara mudah, analisis di atas boleh dirujuk dalam Rajah 4.11.

Secara keseluruhan, kesan yang ketara atau impak gabungan bilangan tugasan karangan dan bilangan pemeriksa terhadap prosedur pemarkahan yang berlainan dari segi kebergantungan skor adalah seperti berikut:

1. Pertambahan bilangan tugasan adalah lebih berkesan berbanding dengan pertambahan bilangan pemeriksa untuk mencapai pekali G yang tinggi merentas semua prosedur pemarkahan.
2. Peningkatan peratusan pekali G yang lebih besar akan diperoleh apabila bilangan tugasan atau bilangan pemeriksa bertambah daripada satu kepada dua terutamanya bagi faset tugasan tanpa mengira prosedur pemarkahan yang digunakan. Namun begitu, gabungan bilangan tugasan dan bilangan pemeriksa akan mengalami pulangan menurun dari segi pekali G apabila bilangan semakin bertambah.
3. Aspek penggunaan bahasa dan mekanis lebih mudah mencecah nilai pekali yang tinggi berbanding dengan aspek kandungan dan organisasi. Manakala kaedah analitik secara relatif lebih mudah mencapai nilai pekali G yang tinggi berbanding dengan kaedah holistik.
4. Aspek penggunaan bahasa dan mekanis yang digabungkan dengan kaedah analitik adalah prosedur pemarkahan yang paling mudah mencapai pekali G yang tinggi (dapat mencapai kriteria .90 apabila pemadanan $\acute{n}_t = 3$, $\acute{n}_r = 3$).

4.2.2 Analisis keputusan reka bentuk separa tersarang $p \times (R:T)$ model gabungan

Analisis dapatan kajian berikut adalah untuk menjawab soalan kajian 3(c) iaitu sejauh manakah impak gabungan bilangan tugasan karangan dan bilangan pemeriksa dengan faset tugasan ditetapkan bagi prosedur pemarkahan yang berlainan terhadap kebergantungan skor berdasarkan kerangka kajian D.

Jadual 4.18 memaparkan keputusan perubahan pekali G untuk kombinasi bilangan tugasan dan pemeriksa yang berbeza dengan tugasan sebagai faset tetap bagi prosedur pemarkahan yang berlainan berdasarkan reka bentuk $p \times (R: T)$ kajian D model gabungan. Dapatan kajian tersebut adalah berdasarkan kombinasi bilangan tugasan karangan dan bilangan pemeriksa yang berlainan yang melibatkan empat buah tugasan (sebagai faset tetap) dan lapan orang pemeriksa. Keputusan bagi operasi pelbagai kajian D berdasarkan reka bentuk tersebut boleh dilihat dalam Lampiran U 1 hingga Lampiran U 4 dan telah dirumuskan dalam Jadual 4.18. Analisis berikut adalah berpandukan keputusan yang dipaparkan dalam Jadual 4.18.

Dapatan kajian menunjukkan bahawa semua prosedur pemarkahan memaparkan pekali G yang lebih tinggi daripada .80 (antara .817 hingga .842) berdasarkan senario kombinasi bilangan tugasan dan pemeriksa tunggal dengan tugasan sebagai faset tetap. Antaranya, nilai pekali bagi kaedah holistik serta aspek penggunaan bahasa dan mekanis adalah hampir sama dengan kaedah analitik serta aspek kandungan dan organisasi (.842 dan .838) manakala pekali bagi kaedah holistik serta aspek kandungan dan organisasi pula adalah hampir sama dengan kaedah analitik serta aspek penggunaan bahasa dan mekanis (.818 dan .817). Namun

Jadual 4.18

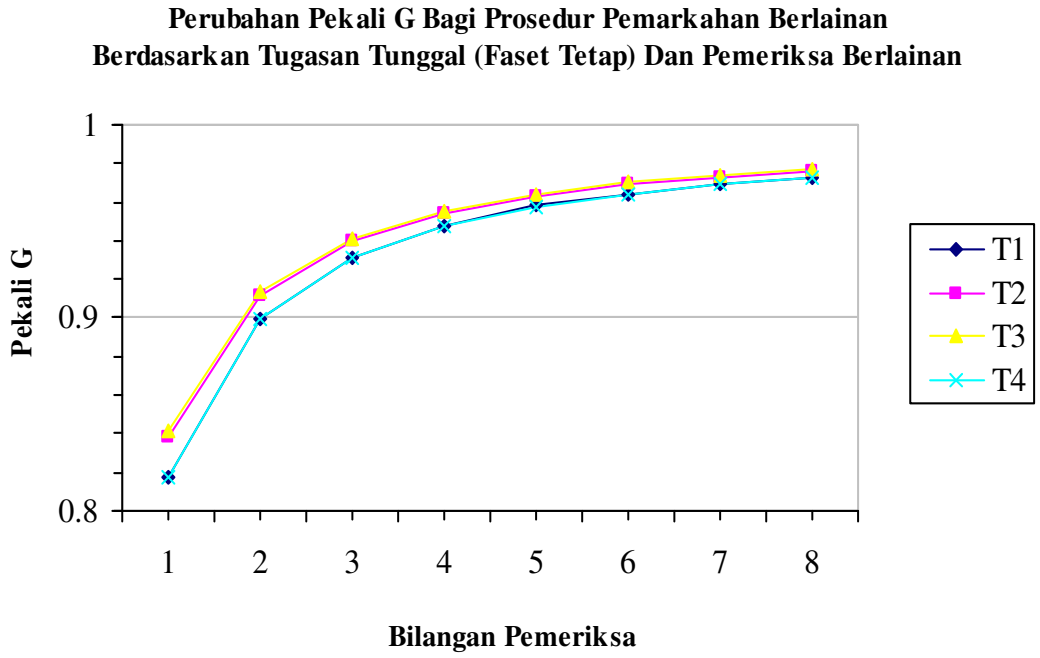
Keputusan Perubahan Pekali G Bagi Bilangan Tugasan Dan Pemeriksa Dengan Tugasan Ditetapkan Di Bawah Prosedur Pemarkahan Yang Berlainan Berdasarkan Reka Bentuk p ×(R: T) Kajian D

Bil Tugasan (Faset Tetap)	Bil Pemeriksa	Aspek Kandungan Dan Organisasi		Aspek Penggunaan Bahasa Dan Mekanis	
		Kaedah Holistik	Kaedah Analitik	Kaedah Holistik	Kaedah Analitik
1	1	.818	.838	.842	.817
	2	.900	.912	.914	.899
	3	.931	.940	.941	.931
	4	.947	.954	.955	.947
	5	.958	.963	.964	.957
	6	.964	.969	.970	.964
	7	.969	.973	.974	.969
	8	.973	.976	.977	.973
2	1	.883	.899	.903	.890
	2	.938	.947	.949	.942
	3	.958	.964	.965	.960
	4	.968	.973	.974	.970
	5	.974	.978	.979	.976
	6	.978	.982	.982	.980
	7	.981	.984	.985	.983
	8	.984	.986	.987	.985
3	1	.914	.926	.930	.921
	2	.955	.962	.964	.959
	3	.969	.974	.975	.972
	4	.977	.981	.982	.979
	5	.981	.984	.985	.983
	6	.985	.987	.988	.986
	7	.987	.989	.989	.988
	8	.988	.990	.991	.989
4	1	.932	.942	.945	.939
	2	.965	.970	.972	.968
	3	.976	.980	.981	.979
	4	.982	.985	.986	.984
	5	.986	.988	.989	.987
	6	.988	.990	.990	.989
	7	.990	.991	.992	.991
	8	.991	.992	.993	.992

begitu, dapatan kajian menunjukkan apabila pemeriksa kedua digunakan, semua prosedur pemarkahan mampu mencapai kriteria pekali .90 (Rajah 4.12 hingga Rajah 4.15). Penggunaan bilangan pemeriksa yang ke-3 dan lebih dengan tugasan tunggal ditetapkan adalah tidak banyak memberi hasil dari segi peningkatan kepersisan pengukuran.

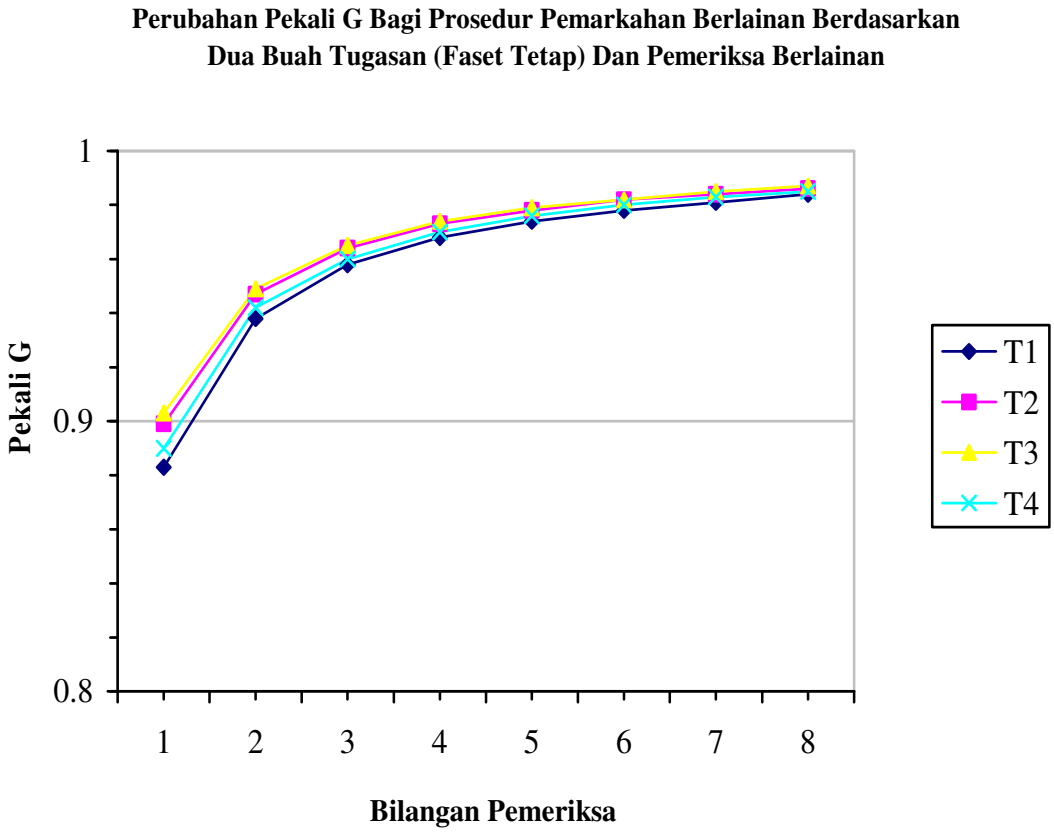
Untuk senario kombinasi bilangan dua buah tugasan dan pemeriksa tunggal ($\acute{n}_t = 2$, $\acute{n}_r = 1$) dengan tugasan sebagai faset tetap, dapatan kajian jelas menunjukkan bahawa kaedah holistik serta aspek penggunaan bahasa dan mekanis (.903) di samping kaedah analitik serta aspek kandungan dan organisasi (.899) mampu menepati kriteria .90 (Rajah 4.13). Prosedur pemarkahan selebihnya mampu mencecah kriteria .90 apabila pemeriksa kedua digunakan. Sekiranya bilangan tugasan karangan untuk reka bentuk asal kajian ini ($\acute{n}_t = 2$, $\acute{n}_r = 3$) dipiawaikan, maka semua prosedur pemarkahan dapat memperoleh pekali G yang tinggi iaitu antara .958 hingga .965. Dapatan kajian menunjukkan penggunaan dua orang pemeriksa sudah memadai untuk menepati kriteria .90 bagi semua prosedur pemarkahan dalam kajian ini. Pertambahan bilangan pemeriksa yang seterusnya tidak banyak menyumbang kepada kejituan pengukuran. Manakala apabila tugasan ditetapkan pada $\acute{n}_t = 3$, semua prosedur pemarkahan dapat mencecah pekali G yang lebih tinggi daripada .90 berdasarkan pemeriksa tunggal sahaja (antara .914 hingga .930). Pekali G yang lebih tinggi lagi akan diperoleh apabila tugasan ditetapkan pada $\acute{n}_t = 4$ dengan pemeriksa tunggal (antara .932 hingga .945). Bagi senario keadaan tugasan yang ditetapkan pada $\acute{n}_t = 2$ dan ke atas, penggunaan pemeriksa yang lebih daripada dua orang adalah tidak membawa sebarang signifikan dari segi kepersisan pengukuran.

Jika diteliti, dapatan kajian menunjukkan perubahan bilangan tugasan dan pemeriksa daripada satu kepada dua dengan tugasan ditetapkan akan membawa kepada peningkatan pekali yang dramatik (Rajah 4.12 hingga Rajah 4.15). Misalnya, tanpa mengira prosedur pemarkahan yang digunakan, apabila bilangan tugasan (faset tetap) berubah daripada satu kepada dua dengan pemeriksa tunggal digunakan, peratusan pekali G akan meningkat antara 7.3% hingga 8.9% manakala pertambahan



Rajah 4.12. Pekali G bagi prosedur pemarkahan berlainan berdasarkan tugasan tunggal dan bilangan pemeriksa berlainan dengan faset tugasan ditetapkan.

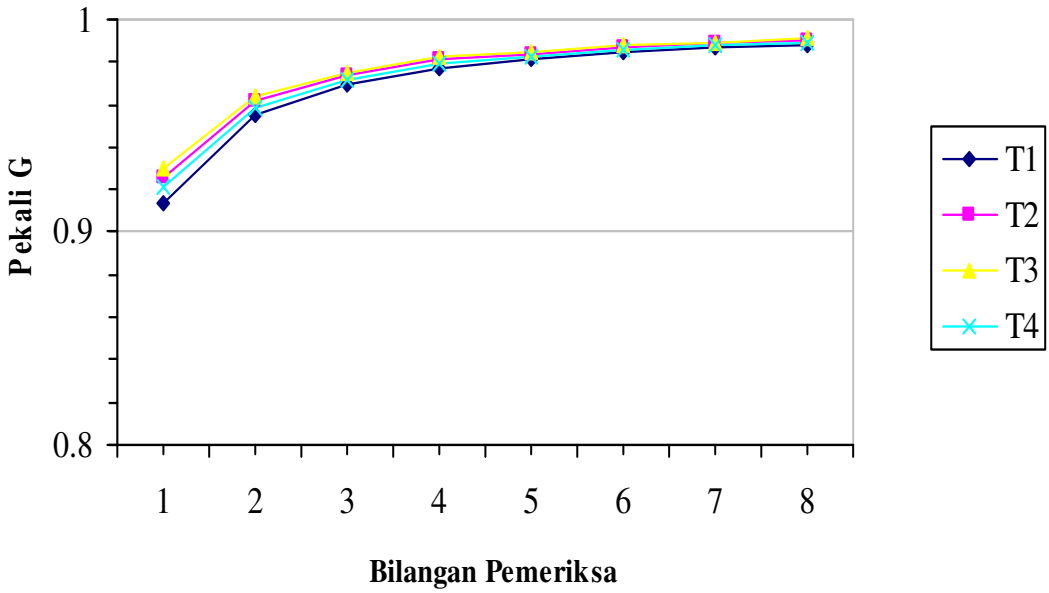
Nota. T1= Kaedah Holistik Serta Aspek Kandungan Dan Organisasi; T2= Kaedah Analitik Serta Aspek Kandungan Dan Organisasi; T3= Kaedah Holistik Serta Aspek Penggunaan Bahasa Dan Mekanis; T4= Kaedah Analitik Serta Aspek Penggunaan Bahasa Dan Mekanis.



Rajah 4.13. Pekali G bagi prosedur pemarkahan berlainan berdasarkan dua buah tugasan dan bilangan pemeriksa berlainan dengan faset tugasan ditetapkan

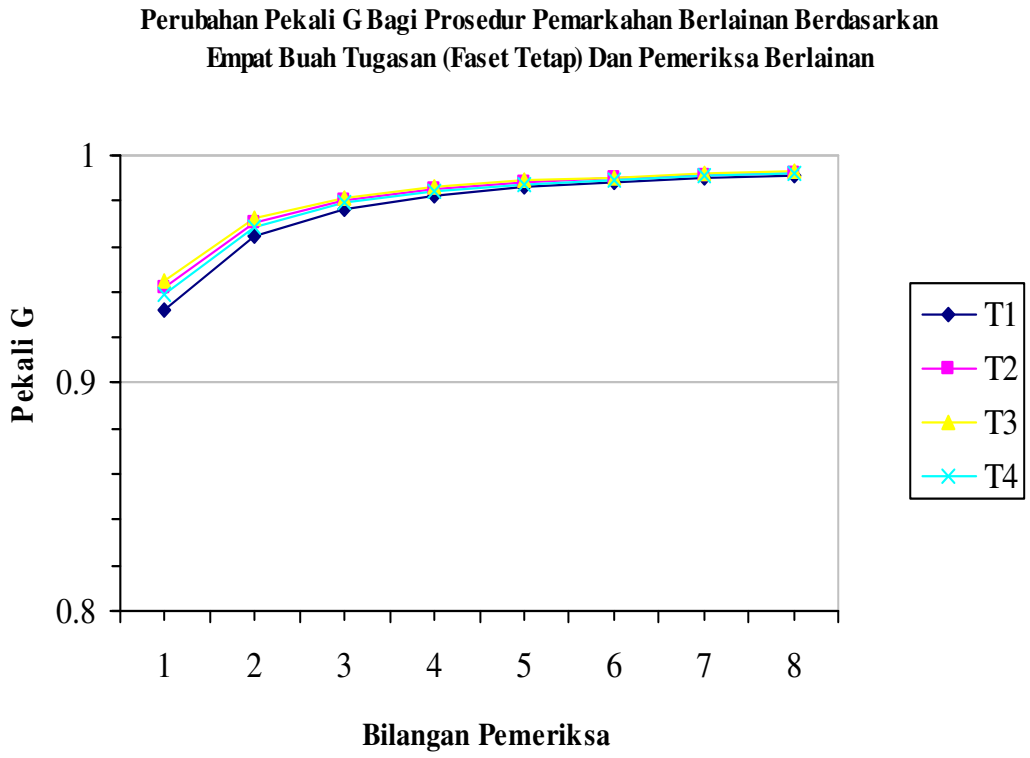
Nota. T1= Kaedah Holistik Serta Aspek Kandungan Dan Organisasi; T2= Kaedah Analitik Serta Aspek Kandungan Dan Organisasi; T3= Kaedah Holistik Serta Aspek Penggunaan Bahasa Dan Mekanis; T4= Kaedah Analitik Serta Aspek Penggunaan Bahasa Dan Mekanis.

Perubahan Pekali G Bagi Prosedur Pemarkahan Berlainan Berdasarkan Tiga Buah Tugasan (Faset Tetap) Dan Pemeriksa Berlainan



Rajah 4.14. Pekali G bagi prosedur pemarkahan berlainan berdasarkan tiga buah tugasan dan bilangan pemeriksa berlainan dengan faset tugasan ditetapkan.

Nota. T1= Kaedah Holistik Serta Aspek Kandungan Dan Organisasi; T2= Kaedah Analitik Serta Aspek Kandungan Dan Organisasi; T3= Kaedah Holistik Serta Aspek Penggunaan Bahasa Dan Mekanis; T4= Kaedah Analitik Serta Aspek Penggunaan Bahasa Dan Mekanis.



Rajah 4.15. Pekali G bagi prosedur pemarkahan berlainan berdasarkan empat buah tugasan dan bilangan pemeriksa berlainan dengan faset tugasan ditetapkan

Nota. T1= Kaedah Holistik Serta Aspek Kandungan Dan Organisasi; T2= Kaedah Analitik Serta Aspek Kandungan Dan Organisasi; T3= Kaedah Holistik Serta Aspek Penggunaan Bahasa Dan Mekanis; T4= Kaedah Analitik Serta Aspek Penggunaan Bahasa Dan Mekanis.

bilangan pemeriksa daripada satu kepada dua dengan tugasan tunggal digunakan akan menyebabkan peningkatan peratusan pekali yang lebih tinggi lagi iaitu antara 8.6% hingga 10.0%. Ini juga bermakna pertambahan bilangan pemeriksa adalah lebih berkesan dalam meningkatkan kebergantungan skor berbanding dengan pertambahan dalam bilangan tugasan (faset tetap). Keputusan ini dapat dijangkakan kerana semesta generalisasi untuk faset tugasan telah dihadkan dan hanya tinggal faset pemeriksa sebagai sumber varian ralat. Pertambahan bilangan tugasan (faset tetap) dan pemeriksa yang seterusnya menampakkan peningkatan pekali yang menurun. Keputusan kajian juga menunjukkan penggunaan bilangan tugasan (faset tetap) atau

bilangan pemeriksa yang semakin meningkat terutamanya tugasan (faset tetap) tidak banyak menyumbang ke arah kepersisian pengukuran.

Jika dianalisis dari segi aspek pemarkahan dan kaedah pemarkahan berdasarkan kombinasi bilangan tugasan (faset tetap) dan bilangan pemeriksa pada kadar yang sama, aspek penggunaan bahasa dan mekanis secara relatif mempunyai sedikit kelebihan berbanding dengan aspek kandungan dan organisasi dari segi senangnya mencecah nilai pekali yang tinggi. Ini dapat dilihat terutamanya bagi pemarkahan kaedah holistik iaitu apabila karangan tunggal (faset tetap) dan dua orang pemeriksa digunakan, pekali G akan mencecah .914. Secara teliti, pekali G bagi prosedur pemarkahan kaedah holistik serta aspek penggunaan bahasa dan mekanis ini akan meningkat dengan bertambahnya bilangan tugasan (faset tetap) dan bilangan pemeriksa. Nilai pekalinya juga sentiasa lebih tinggi berbanding dengan tiga prosedur pemarkahan yang lain berdasarkan bilangan satu hingga empat buah tugasan (faset tetap) dan satu hingga lapan orang pemeriksa. Walaupun prosedur-prosedur yang menggunakan aspek kandungan dan organisasi mampu mencecah pekali G pada tahap .90 dan .91 masing-masing dengan kombinasi bilangan karangan dan pemeriksa yang sama, namun nilai pekali bagi prosedur-prosedur tersebut masih kurang sedikit daripada kaedah holistik serta aspek penggunaan bahasa dan mekanis.

Dari segi kaedah pemarkahan, dapatan kajian telah menunjukkan keputusannya adalah bercampur-campur. Misalnya, kaedah holistik (serta aspek penggunaan bahasa dan mekanis) menunjukkan kebolehan generalisasi yang paling baik antara semua prosedur pemarkahan manakala kaedah holistik serta aspek kandungan dan organisasi pula menunjukkan kebolehan generalisasi yang paling

lemah. Ini bermakna aspek pemarkahan secara relatif merupakan pertunjuk yang lebih baik daripada kaedah pemarkahan dalam mentaksir kemahiran menulis calon apabila faset tugasan dipiawaikan dan faset pemeriksa adalah rawak.

Meninjau dari segi prosedur pemarkahan pula, keempat-empat prosedur pemarkahan tersebut memaparkan dapatan kajian yang tidak jauh berbeza antara satu sama lain. Perbezaan nilai pekali G antara prosedur-prosedur pemarkahan tersebut akan menjadi tidak begitu ketara apabila nilainya dibundarkan menjadi dua tempat perpuluhan. Walaupun dari segi kombinasi bilangan tugasan (faset tetap) dan bilangan pemeriksa berdasarkan pekali G yang tinggi adalah agak sama, namun aspek penggunaan bahasa dan mekanis secara relatif dapat mencapai nilai pekali G yang sedikit lebih tinggi berbanding dengan aspek kandungan dan organisasi.

Dari segi kaedah pemarkahan pula, kedua-dua kaedah pemarkahan menunjukkan keputusan yang bercampur-campur. Namun begitu, kebergantungan skor bagi prosedur pemarkahan yang menggunakan kaedah holistik serta aspek penggunaan bahasa dan mekanis adalah sedikit lebih tinggi daripada kaedah analitik serta aspek kandungan dan organisasi. Bagi kaedah holistik serta aspek kandungan dan organisasi, kebergantungan skornya adalah lebih rendah sedikit berbanding dengan kaedah analitik serta aspek penggunaan bahasa dan mekanis kecuali bagi senario bilangan tugasan tunggal (faset tetap). Untuk tugasan tunggal yang ditetapkan, pekali bagi kaedah analitik serta aspek penggunaan bahasa dan mekanis adalah tinggi sedikit.

4.3 Rumusan Bab

Bab ini telah memaparkan hasil-hasil keputusan yang diperoleh daripada statistik deskriptif dan statistik inferensi. Selain itu, keputusan komponen-komponen varian berdasarkan reka bentuk separa tersarang $p \times (r{:}t)$ model rawak telah dibentangkan dalam analisis kajian G. Manakala dalam kajian D, keputusan mengenai reka bentuk yang sama berdasarkan model rawak dan model gabungan juga dikemukakan.

BAB V

PERBINCANGAN, IMPLIKASI DAN CADANGAN

5.0 Pengenalan

Bab ini merupakan perbincangan bagi keputusan yang telah diperoleh daripada kajian yang telah dijalankan berdasarkan soalan kajian yang ditetapkan. Skop perbincangan keputusan dan dapatan kajian telah dilihat dari segi teori, dapatan kajian yang lalu dan dapatan kajian ini. Bab ini juga membicarakan implikasi kajian dan mengemukakan cadangan bagi memantapkan kajian-kajian akan datang.

5.1 Perbincangan

Tujuan kajian ini adalah untuk meninjau perkaitan antara kesan relatif tugasan karangan dan pemeriksa serta impak gabungan bilangan tugasan karangan dan pemeriksa terhadap prosedur pemarkahan yang berlainan dari segi kebergantungan skor karangan berdasarkan perspektif teori G. Dapatan-dapatan kajian utama adalah seperti berikut:

(a) Faktor tugasan karangan, pemeriksa dan prosedur pemarkahan mempunyai pengaruh yang besar ke atas keberubahan skor;

(b) Kesan tugasan karangan dan pemeriksa adalah bersandar kepada prosedur pemarkahan yang berlainan;

(c) Sumber varian utama dalam mengkaji kemahiran menulis calon adalah disebabkan oleh perbezaan kebolehan di kalangan calon yang ditaksir dengan tugasan karangan yang berlainan;

(d) Generalisasi skor untuk prosedur pemarkahan yang berlainan adalah berbeza;

(e) Peningkatan pekali G yang paling besar berlaku apabila pemeriksa kedua digunakan;

(f) Cara yang paling berkesan dalam memaksimumkan kebergantungan skor adalah menambahkan bilangan tugasan karangan daripada menambahkan bilangan pemeriksa;

(g) Dari segi aspek pemarkahan: aspek penggunaan bahasa dan mekanis nampaknya lebih mudah mencapai kebergantungan skor yang tinggi manakala dari segi kaedah pemarkahan, kaedah analitik nampaknya secara relatif lebih mudah mencapai kebergantungan skor yang tinggi;

(h) Dari segi prosedur pemarkahan: aspek penggunaan bahasa dan mekanis serta kaedah analitik merupakan prosedur pemarkahan yang paling mudah mencecah kebergantungan skor yang tinggi;

(i) Dengan faset tugasan ditetapkan, semua prosedur pemarkahan menunjukkan peningkatan yang mendadak dari segi pekali G.

Semua dapatan kajian ini akan dibincang dengan secara terperinci seterusnya, bersama dengan implikasinya berdasarkan prosedur pemarkahan yang berlainan.

5.1.1 Tugasan karangan

Kajian ini menggunakan dua buah karangan yang berlainan tugasan iaitu karangan berunsur naratif dan pendedahan. Dapatan kajian menunjukkan tugasan karangan yang berlainan mempunyai pengaruh tertentu terhadap prestasi skor karangan calon. Analisis statistik deskriptif menunjukkan bahawa kedua-dua tugasan karangan adalah berbeza dari segi aras kesukuran. Ini dapat dilihat berdasarkan prosedur pemarkahan yang berlainan, skor calon dalam karangan berunsur naratif

pada keseluruhannya adalah lebih tinggi daripada skor karangan berunsur pendedahan. Kajian Scott (1996) ke atas bentuk karangan yang berlainan juga mendapati karangan bentuk pendedahan adalah lebih sukar daripada bentuk naratif dan deskriptif sama ada dalam pentaksiran bahasa pertama atau kedua.

Analisis ujian *Levene* juga menunjukkan bahawa varian bagi kedua-dua tugasan karangan berdasarkan prosedur pemarkahan yang berlainan adalah tidak homogen dan ralat pengukuran bagi karangan berunsur pendedahan cenderung lebih besar berbanding dengan karangan berunsur naratif. Selain itu, analisis ujian *t* menunjukkan perbezaan min skor bagi tugasan karangan yang berlainan adalah signifikan dalam dua prosedur pemarkahan iaitu aspek kandungan dan organisasi serta aspek penggunaan bahasa dan mekanis berdasarkan kaedah holistik. Namun begitu, kesan saiz bagi kedua-dua tugasan karangan berdasarkan prosedur pemarkahan tersebut ke atas skor karangan adalah kecil iaitu masing-masing adalah .049 dan .021.

Walaupun analisis yang seterusnya iaitu kajian G tidak dapat mengungkaikan amaun varian ralat bagi kedua-dua tugasan karangan tersebut secara berasingan, namun sekiranya ditinjau dari segi prosedur pemarkahan yang berlainan, komponen varian untuk interaksi calon dan tugasan $\sigma^2(pt)$ telah menyumbangkan amaun varian yang agak besar daripada jumlah varian adalah jelas. Secara teliti, dalam semua prosedur pemarkahan kecuali prosedur aspek penggunaan bahasa dan mekanis serta kaedah analitik, menunjukkan bahawa komponen varian $\sigma^2(pt)$ merupakan komponen kedua terbesar selepas varian benar untuk calon. Malah komponen varian $\sigma^2(pt)$ bagi prosedur aspek kandungan dan organisasi serta kaedah holistik meliputi kira-kira

separuh daripada varian benar untuk calon (26.7% berbanding dengan 55% daripada varian keseluruhan). Meskipun prosedur aspek penggunaan bahasa dan mekanis serta kaedah analitik adalah penyumbang varian $\sigma^2(pt)$ yang paling kecil antara semua prosedur pemarkahan tetapi amaunnya masih meliputi 15.4% daripada varian keseluruhan dan merupakan komponen varian ketiga terbesar. Ini juga bermakna anggapan bahawa aras kesukaran karangan berunsur pendedahan adalah lebih tinggi daripada karangan berunsur naratif adalah anggapan yang munasabah.

Selain itu, komponen varian $\sigma^2(pt)$ yang agak besar ini juga menunjukkan bahawa susunan kedudukan (*rank order*) calon berdasarkan dua buah karangan yang digunakan dalam kajian ini adalah berbeza. Dapatan ini secara tidak langsung menandakan bahawa sebarang generalisasi tentang kedudukan relatif calon dengan menggunakan salah sebuah karangan tersebut mungkin kurang boleh dipercayai dan akan menghasilkan interpretasi yang berlainan terhadap kemahiran menulis calon. Ini juga bermakna generalisasi tentang kemahiran menulis calon lebih boleh dipercayai sekiranya lebih daripada sebuah buah karangan digunakan.

Bentuk karangan yang berbeza mengkehendaki tahap penglibatan pemikiran dan komitmen penulis yang berlainan (Fairbarin & Winch, 1996). Langer (1984) juga mengungkitkan kesangsian terhadap kewajaran ekadimensi tentang pengetahuan. Beliau telah membezakan antara latar belakang pengetahuan yang dimiliki dengan kebolehan mengorganisasi pengetahuan, dan beranggapan bahawa yang pertama lebih berkesan dalam menghasilkan karangan yang berasaskan penyampaian maklumat (pendedahan) sementara yang kedua pula lebih berkesan

dalam menjana karangan berunsur pembahasan. Jelaslah bahawa karangan berunsur pendedahan lebih bergantung kepada banyaknya pengetahuan yang ada pada penulis.

Kajian ini juga memberi serba sedikit sokongan dari segi empirikal tentang apa yang membentuk kemahiran menulis atau gagasan kemahiran menulis. Struktur pembentukan kemahiran menulis lebih menyerupai satu entiti yang tersusun, proses pemerolehannya adalah berperingkat. Setidak-tidaknya dalam kajian empirikal ini, hipotesis mengenai wujudnya konsep ekadimensi tentang pengetahuan adalah kabur dan kurang meyakinkan. Ini mungkin kerana sifat karangan berunsur pendedahan yang menuntut calon menerangkan dan menjelaskan sesuatu perkara atau maklumat di samping lebih menekankan penyampaian secara tersusun dan logik. Oleh itu, penulis perlu memiliki kebolehan berbahasa yang lebih mantap selain memiliki pengetahuan yang cukup tentang perkara atau maklumat tertentu. Dapatan kajian menunjukkan bahawa bagi pentaksiran aspek kandungan dan organisasi, nilai varian bagi *pt* adalah lebih besar (kaedah analitik, 23.8%; kaedah holistik, 26.7%). Ini mungkin menunjukkan calon masih perlu dilengkapi dengan pengetahuan yang cukup untuk menangani tugasan berunsur pendedahan memandangkan pengalaman dan kebiasaan calon terhadap tugasan tersebut masih terbatas. Ini juga menyebabkan pencapaian calon akan lembab sedikit apabila menjawab tugasan tersebut. Sebaliknya, tugasan berunsur naratif menuntut calon menggunakan teknik cerita yang mudah berdasarkan pengalaman dan daya imaginatif yang sesuai dengan perkembangan kognitif murid. Secara kesimpulan, struktur pengetahuan yang perlu dimiliki oleh calon dalam kedua-dua tugasan tersebut adalah berlainan.

5.1.2 Prosedur pemarkahan

5.1.2.1 Kaedah pemarkahan

Jika dilihat dari segi prosedur pemarkahan, berdasarkan dapatan kajian jelas menunjukkan kaedah pemarkahan yang digunakan boleh mempengaruhi kebolehpercayaan skor karangan (Breland, 1983; Coffman, 1971a & 1971b) dan memberi pengaruh tertentu terhadap varian skor calon. Keputusan kajian ini menunjukkan kaedah analitik secara relatif mempunyai pekali yang lebih tinggi daripada kaedah holistik berdasarkan prosedur pemarkahan yang berlainan. Dapatan kajian lepas juga melaporkan bahawa anggaran pekali kebolehpercayaan pemarkahan kaedah analitik adalah lebih tinggi daripada kaedah holistik (misalnya, Huot, 1990a; Klein et al., 1998). Dari segi teori, ujian yang mempunyai bilangan item dan tugasan yang lebih banyak biasanya mempunyai kebolehpercayaan skor yang lebih tinggi (Allen & Yen, 1979; Linn & Gronlund, 2000). Secara logik, pemarkahan analitik yang mempunyai bilangan subskala yang lebih banyak akan memberi kelebihan dari segi ketekalan pemarkahan secara keseluruhan (lihat Brown & Bailey, 1984; Hamp-Lyons, 1991). Dalam kajian ini, pemeriksa yang menggunakan kaedah analitik mempunyai peluang untuk mentaksir lebih daripada satu aspek pemarkahan. Justeru itu, ralat skor secara purata akan berkurangan dan kebolehpercayaan skor akan meningkat.

Selain itu, dari sudut kognitif psikologi, jika pemeriksa mempunyai skema psikologi terhadap prestasi karangan calon, maka apabila prosedur pemarkahan yang berlainan diberikan akan menyebabkan pemeriksa mengubahsuai skema asalnya berdasarkan skema baru. Sekiranya skema baru bercanggah dengan skema asal dan pemeriksa yang berlainan pula mempunyai pemahaman dan daya penyesuaian yang

berlainan terhadap skema baru, ini akan mempengaruhi pemarkahan pemeriksa. Tambahan pula, kehendak setiap prosedur pemarkahan yang berlainan terhadap daya ingatan dan kebiasaan kognitif pemeriksa juga adalah berlainan. Dalam kajian ini, ralat pengukuran untuk pemarkahan kaedah analitik adalah lebih kecil mungkin kerana kaedah tersebut lebih menepati skema psikologi dan kebiasaan kognitif kebanyakan pemeriksa. Para pemeriksa dalam kajian ini sudah biasa dengan kaedah pemarkahan analitik kerana mereka menggunakan kaedah tersebut untuk mentaksir kemahiran menulis murid di sekolah masing-masing untuk jangka masa yang lama. Selain itu, mereka juga berpengalaman untuk menilai skrip karangan Bahasa Cina UPSR yang sebelum ini menggunakan kaedah analitik. Oleh itu, kaedah pemarkahan analitik mungkin lebih menghampiri skema psikologi dan kebiasaan kognitif pemeriksa (Hayes, Hatch, & Silk, 2000).

5.1.2.2 Aspek pemarkahan

Dapatan kajian ini juga membekalkan serba sedikit panduan untuk mendasari pemilihan aspek pemarkahan bagi pentaksiran kemahiran menulis calon. Ralat pengukuran untuk skor aspek penggunaan bahasa dan mekanis adalah lebih kecil daripada aspek kandungan dan organisasi. Aspek penggunaan bahasa dan mekanis nampaknya mempunyai pekali kebolehpercayaan dan kebolehan generalisasi yang lebih tinggi daripada aspek kandungan dan organisasi. Ini bermakna dalam pentaksiran karangan untuk menilai kemahiran menulis murid, aspek penggunaan bahasa dan mekanis iaitu termasuk penggunaan tanda baca, karakter, kosa kata, tatabahasa dan sebagainya mempunyai ciri diskriminasi yang lebih menonjol daripada aspek kandungan dan organisasi. Secara perbandingan, kajian Schoonen (2005) ke atas pemarkahan empat jenis karangan berdasarkan tajuk dan bentuk

karangan yang berlainan juga menunjukkan bahawa pekali G bagi kandungan dan organisasi adalah paling rendah (.21) apabila kaedah analitik digunakan manakala pekali G bagi penggunaan bahasa adalah lebih memuaskan (.40) apabila kaedah holistik digunakan untuk penilaian sampel karangan. Pemilihan tentang aspek pemarkahan untuk mentaksir kemahiran menulis calon melibatkan isu gagasan kemahiran menulis. Pada hakikatnya, pemilihan aspek pemarkahan adalah berkaitan dengan isu kesahan dalam pengukuran. Bagaimanapun, perkara ini tidak akan disentuh secara mendalam kerana tujuan kajian ini adalah untuk meninjau faset-faset pengukuran yang mempengaruhi ralat skor karangan.

5.1.3 Pemeriksa

Oleh sebab kajian ini menggunakan reka bentuk separuh tersarang dengan pemeriksa tersarang dalam tugasan, maka kesan pemeriksa adalah terbaur dalam interaksi pemeriksa × tugasan. Oleh itu, anggaran tentang kesan pemeriksa adalah tidak dapat diasingkan. Namun begitu, anggaran kesan pemeriksa secara kasar masih boleh dijalankan walaupun terdapat keterbatasan dari segi pentafsiran. Berdasarkan komponen-komponen varian $\sigma^2(r{:}t)$ dan $\sigma^2(pr{:}t)$, kesan pemeriksa untuk kaedah holistik adalah lebih kecil berbanding dengan kesan tugasan manakala kesan tugasan adalah lebih kecil untuk kaedah analitik berbanding dengan kesan pemeriksa. Selain itu, kesan pemeriksa adalah lebih kecil untuk aspek kandungan dan organisasi berbanding dengan kesan tugasan manakala kesan tugasan adalah lebih kecil untuk aspek penggunaan bahasa dan mekanis berbanding dengan kesan pemeriksa. Keputusan ini adalah selari dengan keputusan Schoonen et al. (1997) walaupun kajiannya melibatkan pemeriksa pakar dan pemeriksa biasa manakala kajian ini hanya melibatkan pemeriksa berpengalaman sahaja. Keputusan ini juga

menunjukkan kesan bagi suatu faset dalam pentaksiran karangan, iaitu kesan pemeriksa, adalah bersandar kepada kaedah dan aspek pemarkahan yang berlainan.

5.1.4 Tugasan karangan sebagai faset tetap

Dalam kajian ini, keputusan model gabungan dengan faset tugasan (*T*) ditetapkan dan faset pemeriksa (*R*) adalah rawak memperlihatkan peningkatan pekali G yang dramatik apabila bilangan pemeriksa bertambah (Brennan, 2001, p. 126). Ini kerana dengan faset *T* ditetapkan, kedudukan relatif calon berdasarkan tugasan karangan yang berlainan tidak lagi dianggap sebagai varian ralat dalam pengukuran, sebaliknya komponen varian $\sigma^2(pT)$ akan menyumbang kepada varian skor semesta. Komponen varian $\sigma^2(pR{:}T)$ merupakan satu-satunya varian ralat relatif yang tinggal untuk anggaran pekali G. Oleh itu, varian skor semesta dalam semesta generalisasi dengan faset tetap *T* dan faset rawak *R* adalah adalah lebih besar daripada semesta generalisasi bagi kedua-dua faset *T* dan *R* adalah rawak.

Anggaran pekali G adalah lebih besar apabila tugasan dianggap faset tetap kerana semesta generalisasi dengan faset tetap adalah lebih sempit daripada semesta generalisasi dengan kedua-dua fasetnya rawak. Ini bermakna ralat adalah kecil sekiranya generalisasi dilakukan ke atas semesta yang sempit daripada generalisasi dilakukan ke atas semesta yang lebih luas. Keadaan ini dapat dilihat dalam kajian ini apabila kombinasi tugasan dan pemeriksa tunggal (faset tetap), semua prosedur pemarkahan memaparkan pekali G yang lebih tinggi daripada .80 (antara .817 hingga .842) berbanding dengan apabila kedua-dua faset tersebut adalah rawak (antara .551 hingga .663). Namun, ini tidak bermaksud bahawa semesta yang sempit digalakkan kerana mengehadkan suatu semesta akan membataskan takat generalisasi yang boleh

dilakukan oleh seseorang penyelidik (Brennan, 2001, p.2). Dalam kes ini, apabila tugasan ditetapkan, penyelidik tidak boleh menarik kesimpulan tentang apa yang akan berlaku jika tugasan yang berbeza digunakan. Oleh itu, penetapan atau pemiawaian akan menyempitkan interpretasi pengukuran dan menyekat generalisasi yang akan dibuat.

5.1.5 Kajian literatur dan teori

Reka bentuk eksperimen ini adalah hampir sama dengan reka bentuk eksperimen yang digunakan oleh Schoonen ke atas 89 orang murid sekolah rendah di Belanda yang mengkaji skor karangan bahasa ibunda (2005). Sungguhpun begitu, kajian beliau telah menggunakan empat tugasan karangan iaitu tugasan berunsur pemujukan dan deskriptif serta dua tugasan berunsur pendedahan, 20 orang pemeriksa tersarang dalam empat kumpulan yang berbeza di samping kehendak prosedur pemarkahan adalah agak berbeza dengan kajian ini. Selain itu, kajiannya adalah menggunakan model SEM (*Structural Equation Modeling*) untuk menganggar kebolehan generalisasi skor penulisan. Dapatan kajiannya menunjukkan varian ralat bagi interaksi calon dan tugasan karangan yang berlainan $\sigma^2(pt)$ dan varian ralat reja $\sigma^2(pr,prt,e)$ adalah melebihi varian benar untuk calon $\sigma^2(p)$. Sementara itu, varian ralat bagi kaedah holistik adalah lebih kecil (.40 dan .32) berbanding dengan kaedah analitik (.30 dan .21) berdasarkan aspek pemarkahan yang berlainan, dan varian ralat bagi aspek penggunaan bahasa adalah lebih kecil (.40 dan .30) berbanding dengan aspek kandungan dan organisasi (.32 dan .21) berdasarkan kaedah pemarkahan yang berlainan. Dapatan keputusan Schoonen agak berbeza sedikit daripada dapatan keputusan ini. Walaupun kedua-dua varian ralat dalam kajian ini tidak melebihi varian benar untuk calon $\sigma^2(p)$, namun jumlah kedua-dua varian ralat tersebut juga

agak besar (antara 33.7% hingga 45.0% daripada jumlah varian) terutamanya varian ralat $\sigma^2(pt)$. Manakala kajian ini juga menunjukkan bahawa varian ralat bagi aspek penggunaan bahasa dan mekanis adalah lebih kecil berbanding dengan aspek kandungan dan organisasi. Namun begitu, kajian ini menunjukkan varian ralat bagi kaedah analitik adalah lebih kecil daripada kaedah holistik. Perbezaan kajian ini dengan kajian Schoonen mungkin disebabkan oleh kedua-dua kajian ini menggunakan bahan kajian dan sampel kajian yang berlainan. Selain itu, kajian ini dijalankan dalam keadaan yang lebih terkawal dari segi pemilihan tugasan karangan dan latihan pemeriksa yang lebih terperinci. Perbezaan aspek pemarkahan dalam kedua-dua kajian juga mungkin menyebabkan keputusan yang berlainan. Misalnya dalam penskoran analitik, Schoonen menggunakan panduan penskoran yang ketat untuk menentukan sama ada hadir atau tidak sesuatu proposisi yang berkaitan (untuk kandungan dan organisasi) atau kesilapan bahasa (untuk penggunaan bahasa). Aspek pemarkahan yang ditaksir dalam kajian ini pula berdasarkan aspek-aspek kemahiran menulis yang ditekankan dalam kurikulum.

Dapatan kajian ini pada keseluruhannya menunjukkan kesan tugasan karangan yang berlainan, kesan pemeriksa dan kesan prosedur pemarkahan memang mempunyai pengaruh yang penting terhadap varian skor karangan calon. Faset-faset tersebut telah melemahkan sumbangan kesan objek pengukuran. Ini menyebabkan pertambahan ralat pengukuran bagi keseluruhan pentaksiran karangan, dan seterusnya pekali G yang diperoleh adalah kurang daripada anggaran yang dihasratkan. Meskipun begitu, dapatan kajian ini telah memenuhi hasrat tujuan kajian asal iaitu sama ada interaksi faset-faset berkenaan akhirnya mempengaruhi perubahan pekali G yang dihasilkan serta setakat mana perubahan tersebut

mempengaruhi keputusan kajian yang diperoleh. Jika dilihat dari segi dapatan kajian berdasarkan penggunaan reka bentuk eksperimen $p \times (r: t)$ iaitu reka bentuk kajian yang digunakan dalam kajian ini, dapatan kajian pentaksiran komponen penulisan dalam program *EXPLORE* juga memperoleh pekali G yang kurang memuaskan iaitu .37 (ACT, 1994). Manakala dapatan kajian *Manoa Writing Placement Test* (MWPT) memperoleh nilai pekali G agak lebih baik iaitu kira-kira .63 (Brown & Hudson, 2002). Sementara itu, kajian Schoonen (2005) juga menunjukkan pekali G yang kurang memuaskan (paling tinggi .40 dan paling rendah 0.21) berbanding dengan kajian ini yang mendapat keputusan yang agak baik sedikit (paling tinggi .66 dan paling rendah 0.55). Keputusan pekali kebolehpercayaan yang cenderung rendah ini menunjukkan bahawa pemarkahan subjektif adalah terdiri daripada satu proses kompleks yang dibentuk oleh pelbagai rangkaian yang saling berkaitan yang mempengaruhi antara satu sama lain (Grab & Kaplan, 1996). Keadaan ini menyebabkan kita sukar mengukur kemahiran menulis yang bebas keseluruhannya daripada gangguan tugasan karangan, pemeriksa, prosedur pemarkahan dan pelbagai lagi faset-faset pengukuran yang menyumbang kepada ralat skor karangan. Justeru itu, kita perlu melaksanakan lebih banyak kajian untuk memahami dengan secara mendalam tentang hubungan interaksi antara faset, interaksi faset dan objek pengukuran, dan mungkin juga faktor-faktor lain yang berpotensi menjejaskan kebergantungan skor karangan.

5.2 Implikasi Kajian

Kajian ini merupakan satu analisis ke atas kebergantungan skor karangan yang dijalankan berasaskan kerangka kajian teori G untuk melihat perkaitan kompleks antara kesan tugasan karangan, pemeriksa dan prosedur pemarkahan yang

berlainan terhadap kemahiran menulis dalam karangan Bahasa Cina murid Tahun Enam SJK(C) di salah sebuah daerah di Perak. Berdasarkan dapatan kajian dan petunjuk yang diterima, didapati bahawa kesan tugasan karangan dan pemeriksa adalah bersandar kepada prosedur pemarkahan yang berlainan, di samping itu, cara yang paling berkesan untuk memaksimumkan kebergantungan skor adalah dengan menambahkan bilangan tugasan karangan daripada menambahkan bilangan pemeriksa. Selain itu, kebolehan generalisasi aspek penggunaan bahasa dan mekanis adalah lebih kuat daripada aspek kandungan dan organisasi manakala kebolehan generalisasi kaedah analitik secara relatif adalah lebih kuat daripada kaedah holistik. Sementara itu, impak gabungan bilangan tugasan karangan dan pemeriksa berdasarkan prosedur pemarkahan yang berlainan juga adalah berbeza. Jika dilihat dari perspektif pentaksiran menulis karangan sama ada dijalankan di peringkat sekolah, daerah, negeri atau kebangsaan, kesimpulan ini seharusnya dapat menyatakan beberapa implikasi seperti berikut:

5.2.1 Pentaksiran kemahiran menulis

Memandangkan kesan tugasan karangan dan pemeriksa adalah bersandar kepada prosedur pemarkahan (kaedah dan aspek pemarkahan) yang berlainan dalam mentaksir skor karangan, maka penggunaan dan pemilihan kaedah dan aspek pemarkahan yang bersesuaian dalam pentaksiran karangan akan menjamin keadilan, kebolehpercayaan dan kesahan dalam pemberian skor. Kaedah pemarkahan analitik yang mempunyai lebih daripada satu skor memirip kepada mengesan kelemahan penguasaan aspek-aspek tertentu dalam kemahiran menulis adalah lebih sesuai digunakan untuk murid sekolah rendah. Ini memandangkan penguasaan kemahiran menulis untuk murid peringkat sekolah rendah masih tidak seimbang dari segi

penguasaan aspek-aspek kemahiran menulis tertentu. Justeru itu, skor karangan menerusi kaedah pemarkahan analitik yang bersifat diagnosis adalah mempunyai kebolehpercayaan yang lebih. Dari segi aspek pemarkahan, wajaran yang lebih tinggi haruslah diberi kepada aspek penggunaan bahasa dan mekanis daripada aspek kandungan dan organisasi. Ini kerana pentaksiran karangan dengan aspek penggunaan bahasa dan mekanis mempunyai unsur diskriminasi yang tinggi dalam pentaksiran kemahiran menulis. Untuk murid peringkat sekolah rendah, penguasaan penggunaan bahasa dan mekanis adalah lebih diutamakan kerana ia akan meletakkan asas yang kukuh untuk mencapai tahap kemahiran menulis yang lebih tinggi seperti penyusunan idea dan pengolahan kandungan (Applebee, 2000, p.92).

5.2.2 Bilangan tugasan karangan yang berlainan

Dapatan kajian ini menunjukkan kedudukan relatif calon adalah berbeza berdasarkan tugasan karangan yang berlainan [$\sigma^2(pt)$ yang agak besar]. Dapatan ini secara tidak langsung menandakan bahawa sebarang generalisasi tentang kedudukan relatif calon berdasarkan salah sebuah karangan tersebut akan mempunyai kebolehpercayaan yang rendah dan mungkin menghasilkan interpretasi yang berlainan terhadap kemahiran menulis calon. Ini juga bermakna generalisasi tentang kemahiran menulis calon haruslah berdasarkan lebih daripada sebuah karangan. Breland et al. (1987) menyimpulkan bahawa untuk mencapai tahap kebolehpercahayaan skor seperti ujian aneka pilihan yang piawai iaitu .85 - .95, bilangan karangan yang diperlukan adalah sekurang-kurangnya empat buah dan dua bentuk karangan yang berbeza (lihat juga Anastasi, 1982). Kajian ini pula memerlukan sebanyak tiga buah tugasan karangan dan tiga orang pemeriksa untuk mencecah kriteria .90 yang diingini.

5.3 Cadangan Kajian

Hasil daripada dapatan kajian ini, berikut adalah beberapa cadangan yang boleh diambil dan mungkin dapat memberi serba sedikit panduan terhadap pengajaran dan pembelajaran, serta pentaksiran karangan Bahasa Cina terutamanya pada peringkat sekolah rendah. Selain itu, cadangan untuk kajian akan datang juga dikemukakan.

5.3.1 Pengajaran dan pembelajaran menulis karangan

Pelaksanaan pengajaran dan pembelajaran menulis karangan di SJK(C) perlu memberi penekanan pada aspek penulisan yang berkaitan dengan mengorganisasi dan menyampaikan isi kandungn karangan. Kelemahan pelajar menguasai aspek-aspek tersebut juga berlaku dalam menulis karangan Bahasa Malaysia (Abdul Aziz Abdul Talib, 1993). Penguasaan karakter, kosa kata dan tatabahasa yang mantap tidak dapat dinafikan akan menyediakan asas yang kukuh bagi sesuatu bahasa terutamanya untuk murid sekolah rendah, namun keseimbangan antara kedua-duanya haruslah diberi perhatian yang sewajarnya.

Keputusan kajian juga menunjukkan bahawa penguasaan murid terhadap karangan berunsur cerita atau naratif adalah lebih memuaskan daripada karangan berunsur menerangkan pada keseluruhannya. Oleh itu, untuk meningkatkan kemahiran murid dalam menulis karangan berunsur pendedahan, guru bahasa haruslah merangka strategi tertentu untuk memupuk sikap suka membaca di kalangan murid untuk memantapkan asas pengetahuan mereka. Selain itu, murid perlu diberi peluang untuk menjelaskan atau menerangkan konteks yang konkrit

seperti peristiwa-peristiwa yang berlaku di persekitarannya di samping menekankan penyampaian maklumat dengan tersusun dan logik.

5.3.2 Pentaksiran karangan

Untuk tujuan pentaksiran karangan bagi murid sekolah rendah terutamanya murid Tahun Enam SJK(C), pengkaji mencadangkan penambahan wajaran bagi aspek penggunaan bahasa dan mekanis kerana aspek tersebut menunjukkan ciri diskriminasi dalam penguasaan kemahiran menulis di kalangan murid. Manakala dari segi kaedah pemarkahan, dicadangkan kaedah analitik diberi pertimbangan utama dalam menilai produk karangan calon memandangkan kebolehan generalisasi kaedah tersebut adalah lebih tinggi daripada kaedah holistik.

Selain itu, untuk menjamin prinsip keadilan dan kepersisan skor yang benar-benar mencerminkan kemahiran menulis calon, bilangan tugasan karangan dan pemeriksa yang lebih daripada satu harus digunakan terutamanya pertambahan bilangan tugasan karangan kerana ia lebih berkesan dalam meningkatkan kebolehpercayaan skor. Tambahan pula, untuk memaksimumkan kebolehpercayaan skor, pertambahan bilangan tugasan barangkali adalah lebih efisien dari segi kos daripada pertambahan bilangan pemarkahan untuk setiap tugasan (Lee, Kantor, & Mollaun, 2002).

5.3.3 Kajian akan datang

Meskipun teori G merupakan lanjutan daripada teori ujian klasik, tetapi teori ini memiliki teknik statistik yang lebih kuat dan berupaya mengungkai dan menganggar pelbagai sumber varian secara serentak. Ciri ini membolehkan pengkaji

menganalisis komposisi komponen varian dan seterusnya dapat mengkaji kebolehpercayaan skor dengan berkesan. Namun begitu, berikutan dengan kajian yang lebih mendalam tentang isu ralat pengukuran dan kebolehpercayaan skor, sudah pasti akan melibatkan faset kajian yang lebih kompleks, ini menyebabkan pendekatan kajian *univariate* teori G sukar memenuhi permintaan tersebut. Maka teknik statistik pendekatan *multivariate* teori G yang lebih kuat dari segi menganalisis faset kajian yang lebih kompleks adalah diperlukan, penganggarannya juga lebih mudah dan cepat. Selain itu, untuk kajian mengenai isu ralat pengukuran dalam pemarkahan subjektif pada masa akan datang, lebih banyak bilangan faset ralat yang berpotensi boleh ditambahkan ke dalam prosedur pengukuran agar lebih memanfaatkan kajian yang meninjau ralat pengukuran dan kebergantungan skor dalam pentaksiran menulis karangan.

Kekuatan fungsi prosedur ULS dalam *structural equation modeling* (SEM) adalah bersamaan dengan pendekatan ANOVA. Namun, prosedur ini bukan sahaja berupaya menangani reka bentuk tak seimbang dengan mudah, malah berupaya mengungkaikan nilai varian bagi setiap paras dalam faset pengukuran dengan selanjutnya. Misalnya, anggaran sumbangan varian bagi dua paras tugasan karangan dalam reka bentuk kajian ini iaitu karangan berunsur naratif dan pendedahan boleh diungkai dengan selanjutnya. Selain itu, SEM juga dapat menganggar sumbangan varian bagi setiap pemeriksa yang terlibat. Pendekatan ini amat memanfaatkan kajian yang melibatkan pemarkahan subjektif.

Sungguhpun kajian tentang ralat dalam pemarkahan subjektif membolehkan kita memahami isu kebolehpercayaan skor dan menyediakan asas dalam menangani

isu kebolehpercayaan skor yang berkaitan. Namun, kajian tentang pentaksiran memang tidak dapat mengelak daripada meninjau isu kesahan skor iaitu cara yang sesuai dan sah untuk mentaksir kebolehan sebenar kemahiran menulis calon. Ini melibatkan usaha untuk meninjau dengan secara mendalam teras persoalan tentang gagasan kemahiran menulis.

Berdasarkan reka bentuk kajian $p \times (r{:}t)$ dalam kajian ini, nilai pekali G yang dihasilkan mungkin adalah lebih tinggi. Ini kerana prosedur pemarkahan dalam kajian ini dirawat sebagai faset tetap yang tersembunyi (*hidden fixed facet*) dan varian yang berkaitan dengannya dikira sebagai sebahagian daripada varian skor semesta (Brennan, 2001; Lee & Kantor, 2005). Oleh itu, dicadangkan kajian akan datang meninjau kemungkinan ralat pengukuran yang disumbangkan oleh prosedur pemarkahan dengan menjadikannya sebagai faset rawak.

Dalam kajian ini, keadaan (*occasions*) yang berkaitan dengan pentadbiran tugasan karangan tidak diambil kira sebagai faset pengukuran dalam reka bentuk kajian. Namun, ada bukti tertentu dalam kajian lepas yang menunjukkan bahawa kadang-kadang nilai $\sigma^2(pt)$ yang secara relatif lebih besar lebih sesuai dicirikan sebagai $\sigma^2(pto)$ iaitu interaksi calon × tugasan × keadaan (Ruiz-primo et al., 1993; Webb et al., 2000). Malah terdapat bukti menunjukkan bahawa faset keadaan pemarkahan menyumbang secara signifikan kepada varian ralat (Wainer, 1993). Memandangkan faset keadaan mungkin merupakan suatu faset yang relevan dan berpotensi yang boleh menyumbang kepada varian ralat yang agak besar, oleh itu, adalah disyorkan bahawa kajian akan datang cuba menyelidiki faset peristiwa yang melibatkan keadaan pengujian (*testing occasions*) dan keadaan pemarkahan (*rating*

occasions) jika didapati bersesuaian. Keadaan pengujian mungkin melibatkan pentadbiran ujian pada tempat atau masa yang berbeza manakala keadaan pemarkahan pula adalah seperti pemarkahan skrip karangan calon secara berpusat (*centralized marking*) atau dengan kaedah lain.

Dalam banyak keadaan pentaksiran karangan sebagai pentaksiran prestasi dengan menggunakan teori G, keputusan juga sesuai dibuat berdasarkan kumpulan yang padu, misalnya kelas (lihat Kane & Brennan, 1977, p.271) dan sekolah (lihat Candell & Ercikan, 1994, p. 269; Shavelson, Baxter, & Gao, 1993, p. 225) selain keputusan mengenai individu. Dalam kes kajian ini, sekiranya tujuan kajian adalah untuk menilai prestasi pencapaian sekolah atau kelas, maka sekolah atau kelas boleh dijadikan sebagai objek pengukuran dan calon adalah tersarang dalam objek pengukuran. Reka bentuk yang mungkin boleh digunakan adalah $(p{:}s) \times t$ atau $(p{:}k) \times t$ iaitu calon (p) adalah tersarang dalam kumpulan sekolah (s) atau kelas (k) dan semua calon akan menjawab tugasan (t) yang berlainan manakala kedua-dua faset dalam semesta iaitu calon dan tugasan diandaikan sebagai rawak. Dengan itu, adalah dicadangkan bahawa penyelidik akan datang boleh berusaha menerokai kebolehpercayaan skor karangan dari perspektif kumpulan sekolah dan kelas dengan menggunakan teori G.

Tujuan kajian ini adalah untuk meninjau kebergantungan skor dalam pentaksiran karangan Bahasa Cina untuk murid sekolah rendah Tahun Enam di SJK(C). Menerusi analisis sumber-sumber varian yang terdapat dalam kajian ini, didapati bahawa skor karangan bukan sahaja dipengaruhi oleh kesan pemeriksa dan tugasan karangan malah kedua-duanya bersandar kepada prosedur pemarkahan yang

berlainan. Ini serba sedikit merupakan panduan dan rujukan asas kepada pihak-pihak yang berkepentingan seperti guru bahasa, pembina ujian, pakar pentaksiran dan badan peperiksaan agar berikhtiar meminimumkan ralat pengukuran dalam pentaksiran karangan. Oleh itu, pengkaji berharap kajian yang serupa boleh dilaksanakan ke atas karangan Bahasa Malaysia, Bahasa Inggeris dan bahasa-bahasa lain di negara ini berdasarkan kerangka kajian teori G.

5.4 Rumusan Bab

Kajian ini menggunakan teori G untuk menganalisis skor karangan berdasarkan hubungan kompleks antara kesan tugasan karangan dan pemeriksa dengan prosedur pemarkahan (aspek pemarkahan dan kaedah pemarkahan) yang berlainan. Dapatan kajian menunjukkan kesan tugasan karangan dan pemeriksa serta prosedur pemarkahan mempunyai pengaruh yang agak besar ke atas keberubahan skor karangan. Ini dapat dilihat melalui prosedur pemarkahan tentang aspek kandungan dan organisasi serta kaedah holistik, pengaruhnya adalah hampir sama dengan varian benar objek pengukuran iaitu kemahiran menulis calon. Dalam prosedur pemarkahan tersebut, varian ralat merupakan 45% daripada varian keseluruhan. Malah kekuatan kebolehan generalisasi terhadap skor karangan bagi prosedur pemarkahan yang berlainan juga adalah berbeza untuk tujuan membuat keputusan relatif yang boleh dipercayai. Berdasarkan aspek pemarkahan, aspek penggunaan bahasa dan mekanis nampaknya lebih mudah mencapai pekali G yang tinggi berbanding dengan aspek kandungan dan organisasi. Manakala berdasarkan kaedah pemarkahan, kaedah analitik nampaknya secara relatif lebih mudah mencapai pekali G yang tinggi berbanding dengan kaedah analitik. Berdasarkan prosedur pemarkahan, aspek penggunaan bahasa dan mekanis serta kaedah analitik merupakan

prosedur pemarkahan yang paling mudah mencecah pekali G yang tinggi manakala aspek kandungan dan organisasi serta kaedah holistik pula adalah prosedur pemarkahan yang paling sukar dalam mencecah pekali G yang tinggi.

Rujukan

Abdul Aziz Abdul Talib. (1985). Menilai kemahiran pelajar menulis karangan: Satu tinjauan pemarkahan. *Jurnal Dewan Bahasa, Sept.* Kuala Lumpur: Dewan Bahasa dan Pustaka.

Abdul Aziz Abdul Talib. (1993). *Menguji kemahiran bahasa: Prinsip, teknik dan contoh*. Kuala Lumpur: Dewan Bahasa dan Pustaka.

Abdul Shukor Shaari. (2001). Penulisan Karangan: Beberapa proses yang harus dilalui oleh pelajar. *Dewan Bahasa, Jilid 1*(2)*, Feb.* Kuala Lumpur: Dewan Bahasa dan Pustaka.

Aiken, L. R., & Groth-Marnat, G. (2006). *Psychological testing and assessment* (12th ed.). Boston: Pearson Education Group, Inc.

Alderson, J. C., & Banerjee, J. (2002). Language testing and assessment (Part 2). *Language Teaching, 35*(1), 79-113.

Allen, M. J., & Yen, W. M. (1979). *Introduction to measurement theory*. Belmont, CA: Wadsworth.

American College Testing. (1994). *EXPLORE technical manual*. Iowa City, IA: American College Testing.

American Educational Research Association (AERA), American Psychological Association (APA), and the National Council on Measurement in Education (NCME). (1985). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.

Anastasi, A. (1982). *Psychological testing* (5th ed.). New York: Macmillan.

Applebee, A, N. (2000). Alternative models of writing development. In R. Indrisano & J. R. Squire (Eds.), *Perspective on writing: Research, theory, and practice* (pp. 90–110). Newark, Delaware: International Reading Association, Inc.

Applebee, A. N., Langer, J. A., & Mullis, I. V. S. (1986). *Writing: Trends across the decade, 1974-1984* (National Assessment of Educational Progress Rep. No. 15-W-01). Princeton, NJ: Educational Testing Service.

Arshad Abd. Samad. (2004). *Essential of language testing for Malaysian teachers*. Serdang: Universiti putra Malaysia Press.

Ary, D., Jacob, L. C., Razavieh, A. & Sorensen, C. (2006). *Introduction to research in education* (7th ed.). Belmont, CA: Thomson Wadsworth.

Awang Sariyan. (1987). Aspek bahasa dalam peperiksaan: tinjauan berdasarkan karangan pelajar. *Jurnal Dewan Bahasa*, *Sept*. Kuala Lumpur: Dewan Bahasa dan Pustaka.

Azman Wan Chik. (1994). *Pengujian bahasa: Kes bahasa Melayu* (edisi kedua). Kuala Lumpur: Dewan Bahasa dan Pustaka.

Bygate, M., Skehan, P., & Swain, M. (2001). *Researching pedagogic tasks: Second language learning, teaching and testing*. Harlow: Pearson Education Limited.

Bacha, N. (2001). Writing evaluation: what can analytic versus holistic essay scoring tell us? *System, 29*(3), 371-383.

Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford: Oxford University Press.

Bachman, L. F., Lynch, B. K., & Mason, M. (1995). Investigating variability in tasks and rater judgements in a performance test of foreign language speaking. *Language Testing, 12*(2), 238-257.

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice*. Shanghai: Shanghai Foreign Language Educational Press.

Bahaman Abu Samad, & Turiman Suandi. (1999). *Statistics for social science research: With computer application*. Kuala Lumpur: JJ Print & Copy.

Baker, E., Abedi, J., Linn, R., & Niemi, D. (1996). Dimensionality and generalizability of domain-independent performance assessment. *Journal of Educational Research, 89*(4), 197-205.

Barnwell, D. (1989). 'Naive' native speakers and judgements of oral proficiency in Spanish. *Language Testing, 6*(2), 152-163.

Bauer, B. A. (1981). *A study of the reliabilities and cost-efficiencies of three methods of assessment for writing ability*. (ERIC Document Reproduction Service No. ED216357)

Benton, S. L., Corkill, A. J., Sharp, J. M., Downey, R. G., & Khramtsova, I. (1995). Knowledge, interest, and narrative writing. *Journal of Educational Psychology, 87*(1), 66-79.

Bhasah Abu Bakar. (2003). *Asas pengukuran bilik darjah*. Tanjong Malim: Quantum Books.

Bock, M. (1998). Teaching grammar in context. In S. Angelil-Carter (Ed.). *Access to success: Literacy in academic contexts* (pp.53-65). Cape Town: University of Cape Town Press.

Bolus, R. E., Hinofotis, F. B., & Bailey, K. M. (1982). An introduction to Generalizability Theory in second language research. *Language Learning, 32*(2), 245-258.

Boodoo, G. M., & Garlinghouse, P. (1983). Use of the essay examination to investigate the writing skills of undergraduate education majors. *Educational and Psychological Measurement, 43*(4), 1005-1014.

Breland, H.M. (1983). *The direct assessment of writing skill: A measurement review*. (College Board Rep. No. 83-6). New York: College Entrance Examination Board.

Breland, H. M., Bridgeman, B., & Fowles, M. E. (1999). *Writing assessment in admission to higher education: Review and framework*. (College Board Report No. 99-3). New York: College Entrance Examination Board.

Breland, H. M., Camp, R., Johns, R. J. Morris, M.M., & Rock, D. A. (1987). *Assessing writing skill*. (Research Monograph No. 11). New York: College Entrance Examination Board. (ERIC Document Reproduction Service No. ED286920)

Breland, H. M., Lee, Y. W., & Najarian, M., & Muraki, E. (2004). *An analysis of TOEFL-CBT writing prompt difficulty and comparability for different gender groups*. TOEFL Research Report No. 76. Princeton, NJ: Educational Testing Service.

Brennan, R. L. (1983). *Elements of generalizability theory* [M]. Iowa City, IA: American College Testing.

Brennan, R. L. (1992). *Elements of generalizability theory* [M] (rev. ed.). Iowa City, IA: American College Testing.

Brennan, R. L. (1996). Generalizability of performance assessments. In G. W. Philips (Ed.), *Technical issues in large-scale performance assessment* (pp. 19-58). Washington, DC: U.S. Department of Education, Office of Educational Research and Improvement.

Brennan, R. L. (1998). Raw- score conditional standard errors of measurement in generalizability theory. *Applied Psychological Measurement, 22(*4), 307-331.

Brennan, R. L. (2000). Performance assessments from the perspective of generalizability theory. *Applied Psychological Measurement, 24*(4), 339-353.

Brennan, R. L. (2001). *Generalizability theory*. New York: Springer-Verlag.

Brennan, R. L., Gao, X., Colton, D. A. (1998). Generalizability analyses of work keys listening and writing tests. *Educational and Psychological Measurement, 55*(2), 157-176.

Britton, J., Burgess, T., Martin, N., McLeod, A., & Rosen, H. (1975). *The development of writing abilities.* London: Macmillan Education.

Brossell, G. (1983). Rhetorical specification in essay examination topics. *College English, 45*(2), 165-173.

Brossell, G., & Ash, B. H. (1984). An experiment with the wording of essay topics. *College Composition and Communication, 35*(4),423-425.

Brown, H. D. (2004). *Language assessment: Principles and classroom practices*. NY: Pearson Education, Inc.

Brown, J. D. (2005). *Testing in language programs: A comprehensive guide to English language assessment*. Upper Saddle River, NJ: Prentice Hall Regents.

Brown, J. D., & Bailey, K. M. (1984). A categorical instrument for scoring second language writing skills. *Language Learning, 34*(4), 21-42.

Brown, J. D. (2007). Multiple views of L1 writing score reliability. *Second Language Studies, 25*(2), 1-31.

Brown, J. D. & Hudson, T. (2002). Criterion-referenced language testing. Cambridge: Cambridge University Press.

Brown, J. D., Hilgers, Th., & Marsella, J. (1991). Essay prompts and topics: Minimizing the effect of mean differences. *Writing Communication, 8*(4), 533-556.

Bunch, M. B., Littlefair, W. (1988). *Total score reliability in large-scale writing assessment*. Paper presented at the Conference of the Education Commission of the States / Colorado Department of Education Assessment (Boulder, CO, June 1988). (ERIC Document Reproduction Service No. ED310149)

Bygate, M., Skehan, P., & Swain, M. (2001). *Researching pedagogic tasks: Second language learning, teaching and testing*. Harlow: Pearson Education Limited.

Candell, G. L., & Ercikan, K. (1994). On the generalizability of school-level performance assessment scores. *International Journal of Educational Research, 21*, 267-278.

Cantor, N. K., & Hoover, H. D. (1986). *The reliability and validity of writing assessment: an investigation of rater, prompt within mode, and prompt between mode sources of error*. Paper presented at the Annual Meeting of the American Educational Research Association (70th, San Francisco, CA, April 16-20, 1986). (ERIC Document Reproduction Service No. ED269455)

Carr, N. T. (2000). A comparison of the effects of analytic and holistic composition in the context of composition tests. *Issues in Applied Linguistics, 11*(2), 207-241.

Carroll, J. B. (1993). *Human cognitive abilities*. New York: Cambridge University Press.

Cason, G. J., & Cason, C. L. (1984). A deterministic theory of clinical performance rating: Promising early results. *Evaluation and the Health Professions, 7*(2), 221-247.

Chatterji, M. (2003). *Designing and using tools for educational assessment*. Boston: Pearson Educational, Inc.

Chen, E., Niemi, D., Wang, J., Wang, H., & Mirocha, J. (2007). *Examining the generalizability of direct writing assessment tasks*. (CSE Technical Report 718). Los Angeles, CA: National Center for Research on Evaluation, Standards, and Student Testing (CRESST). (ERIC Document Reproduction Service No. ED498482)

Chi, E. (2001). Comparing holistic and analytic scoring for performance assessment with many-facet Rasch measurement. *Journal of Applied Measurement, 2*(4), 379-388.

Chua, Y. P. (2006). *Kaedah dan statistik penyelidikan: Kaedah penyelidikan* (Buku 1). Kuala Lumpur: McGraw-Hill (Malaysia) Sdn. Bhd.

Coakes, S. J. (2005). *SPSS: Analysis without anguish: Version 12.0 for Windows*. Milton: John Wiley & Sons Australia, Ltd.

Coffman, W. E. (1966). On the validity of essay tests of achievement. *Journal of Educational Measurement, 3*(2), 151-156.

Coffman, W. E. (1971a). Essay examinations. In R. L. Thorndike (Ed.), *Educational Measurement* (2nd ed., pp. 271-302). Washington, DC: American Council on Education.

Coffman, W. E. (1971b). On the reliability of ratings of essay examinations in English. *Research in the Testing of English, 5*(1), 24-37.

Cohen. A. D. (1994). *Assessing language ability in the classroom* (2nd ed.). Boston: Heinle & Heinle Publishers.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Cohen, L., Manion, L., & Morrison, K. (2000). *Research Methods in Education* (5th ed.). London, New York: Routledge Falmer.

Congdon, P. J., & McQueen, J. (2000). The stability of rater severity in largr-scale assessment programs. *Journal of Educational Measurement, 37*(2), 163-178.

Cooper, C. R. (1977). Holistic evaluation of writing. In Cooper, C. R., & Odell, L. (Eds.), *Evaluating writing: Describing, measuring, judging* (pp. 3-32). Urbana, IL: National Council of teachers of English. (ERIC Document Reproduction Service No. ED143020)

Cooper, P. L. (1984). *The assessment of writing ability: A review of research* (GRE Board Research Report GREB No. 82-15R; ETS Research Report 84-12). Princeton, NJ: Educational Testing Service.

Crehan, K. D. (1997). *A discussion of analytic scoring for writing performance assessments*. Paper presented at the Annual Meeting of the Arizona Educational Research Association (Phoenix, AZ, October 1997). (ERIC Document Reproduction Service No. ED414336)

Crick, J.E., & Brennan, R.L. (1983). *Manual for GENOVA: A generalized analysis of variance system*. ACT Technical Bulletin No.43. Iowa City, IA : The American College Testing Program.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. Belmont, CA: Wadsworth Group.

Cronbach, L. J., Rajaratnam, N., & Gleser, G. C. (1963). Theory of generalizability: A liberalization of reliability theory (Part 2). *British Journal of Statistical Psychology*, *16*, 137-163.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles*. New York: Wiley.

Crowhurst, M. (1980). Syntactic complexity and teachers' quality ratings of narrations and arguments. *Research in the Teaching of English, 14*(3), 223-231.

Crowley, S. L., Thompson, B., & Worchel, F. (1994). The children's depression inventory: A comparison of generalizability and classical test theory analysis. *Educational and Psychological Measurement, 54*(3), 705-713.

Cumming, A. (1990). Expertise in evaluating second language compositions. *Language Testing, 7*(1), 31-51.

Cumming, A., Kantor, R., & Power, D. E. (2002). Decision making while rating ESL/EFL writing tasks: A descriptive framework. *The Modern Language Journal, 86*(1), 67-96.

Cumming, A., Kantor, R., Powers, D. E., Santos, T., & Taylor, C. (2000). *TOEFL 2000 writing framework: A working paper*. (TOEFL MS-18). Princeton, NJ: Educational Testing Service.

Dai, H. Q., Zhang, F., & Chen, X. F. (1999). *Xinli jiaoyue celiang* [Pengukuran dalam pendidikan dan psikologi]. Guang Zhou: Jinan Daxue Chubanshe.

Davies, A., Brown, A., Elder, C., Hill, K., Lumley, T., & McNamara, T. (2002). *Dictionary of language testing*. Beijing: Foreign Language Teaching and Research Press.

Diederich, P. B., French, J. W., & Carlton, S. T. (1961). *Factors in the judgements of writing ability*. (Research Bulletin 61-15). Princeton, NJ: Educational Testing Service.

Dunbar, S. B., Koretz, D. M., & Hoover, H. D. (1991). Quality control in the development and use of performance assessments. *Applied Measurement in Education, 4*(4), 289-303.

Eckes, T. (2008). Rater types in writing performance assessments: A classification approach to rater variability. *Language testing, 25*(2), 155-185.

Elbow, P. (1999). Individualism and the teaching of writing: Response to Vai Ramanathan and Dwight Atkinson. *Journal of Second Language Writing, 8*(3), 327-338.

Elliot, N., Plata, M., & Zelhart, P. (1990). *A program development handbook for the holistic assessment of writing*. Lanham, MD: University Press of America, Inc. (ERIC Document Reproduction Service No. ED381808)

Fairbarin, G. J., & Winch, C. (1996). *Reading, writing and reasoning: A guide for students* (2nd ed.). Buckingham: Open University Press.

Feldt, L. S., & Brennan, R. L. (1989). Reliability. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 105-146). New York: American Council on Education /Macmillan.

Ferrara, S. (1993). *Generalizability Theory and Scaling: Their roles in writing assessment and implications for performance assessments in other content areas*. Paper presented at the annual meeting of the National Council on Measurement in Education.

Flower, L. (1994). *The construction on negotiated meaning: A cognitive theory of writing*. Carbondale: Southern Illinois University Press.

Foster, P., & Skehan, P. (1996). The influence of planning and task type on second language performance. *Studies in Second Language Acquisition, 16*(3), 299-323.

Fraenkel, J. R., & Wallen, N. E. (2007). *How to design and evaluate research in education* (6th ed.). New York: McGraw-Hill Companies, Inc.

Franken, M., & Haslett, S. (2002). When and why talking can make writing harder. In S. Ransdell & M.L. Barbier (Eds.), *New directions for research in L2 writing.* Volume II (pp. 208-229). Amsterdam: Kluwer.

Freedman, S. W. (1979). How characteristics of student essays influence teachers' evaluations. *Journal of Educational Psychology, 71*(3), 328-338.

Freedman, A., & Pringle, I. (1984). Why students can't write arguments. *English in Education, 18*(2), 73-84.

Gay, L. R. (2003). *Educational research: Competencies for analysis and applications* (7th ed.). Upper Saddle, New Jersey: Pearson Education, Inc.

Godshalk, F. I., Swineford, F., & Coffman, W. E. (1966). *The measurement of writing ability*. New York: College Entrance Examination Board.

Gabrielson, S., Gordon, B., & Engelhard, G. (1995). The effects of task choice on the qualiti of writing obtained in a statewide assessment. *Applied Measurement in Eduacation, 8*(4), 273-290.

Grab, W., & Kaplan, R. B. (1996). *Theory and practice of writing: An applied linguistic perspective*. London: Addison Wesley Longman Limited.

Gleser, G. C., Cronbach, L. J., & Rajaratnam, N. S. (1965). Generalizability of scores influenced by multiple sources of variance. *Psychometrika, 30*(4), 395-418.

Greatorex, J., & Irenka Suto, W. M. (2006). *An empirical exploration of human judgement in the marking of school examinations*. Paper presented at the International Association for Educational Assessment Conference. (Singapore, 21st to 26th May 2006)

Greenberg, K. (1981). *The effects of variations in essay questions on the writing performance of CUNY freshmen.* New York: Instructional Resources Center City University of New York. (ERIC Document Reproduction Service No. ED236266)

Gronlund, N. E. (2003). *Assessment of student achievement* (7th ed.). Boston: Pearson Education, Inc.

Hamp-Lyons, L. (1990). Second language writing: Assessment issues. In B. Kroll (Ed.), *Second language writing: Research insights for the classroom* (pp. 69-87). Cambridge: Cambridge University Press.

Hamp-Lyons, L. (1991a). Reconstructing 'academic writing proficiency'. In L. Hamp-Lyons (Ed.), *Assessing second language writing in academic contexts*. Norwood, NJ: Ablex Publishing Corporation.

Hamp-Lyons, L. (1991b). Scoring procedures for ESL contexts. In L. Hamp-Lyons (Ed.), *Assessing second language writing in academic contexts*. Norwood, NJ: Ablex Publishing Corporation.

Hamp-Lyons, L. (2001). Fourth generation writing assessment. In T. Silva & P. K. Matsuda (Eds.), *On second-language writing* (pp. 117-128). Mahwah, NJ: Lawrence Erlbaum.

Hamp-Lyons, L. (2002). The scope of writing assessment. *Assessing Writing, 8*, 5-16.

Hamp-Lyons, L., & Kroll, B. (1997). *TOEFL 2000 – Writing: Composition, community and assessment* (TOEFL Monograph Series MS-5). Princeton: Educational Testing Service.

Hamp-Lyons, L., & Prochnow, S. (1994). Examining expert judgments of task difficulty on essay tests. *Journal of Second Language Writing, 3*(1), 49-68.

Hasuria Omar. (1999). Cabaran literasi: Keperluan sebenar dalam pemindahan maklumat. Dalam Ambigapathy Pandian (Ed.), *Literasi dalam pendidikan: Perubahan dan cabaran* (pp. 97-107). Kuala Lumpur: Pusat Pembangunan dan Pendidikan Komuniti (CEDC) Sdn. Bhd.

Hayes, J. R., Hatch, J. A., & Silk, C. M. (2000). Does holistic assessment predict writing performances? Estimating the consistency of student performance on holistically scored writing assignments. *Written Communication, 17*(1), 3-26.

Heaton, J. B. (1979). *Writing English language tests*. London: Longman Group UK Ltd.

Heaton, J. B. (1990). *Classroom testing*. London: Longman Group UK Ltd.

Henning, G. (1996). Accounting for nonsystematic error in performance ratings. *Language Testing, 13*(1), 53-61.

Hidi, S., & Anderson, V. (1992). Situational interest and its impact on reading and expository writing. In K. A. Renninger, S. Hidi, & A. Krapp (Eds.), *The role of interest in learning and development* (pp. 215–238). Hillsdale, NJ: Erlbaum.

Hidi, S., & McLaren, J. (1990). The effect of topic and theme interestingness on the production of school expositions. In H. Mandl, E. De Corte, N. Bennett, & H. F. Friedrich (Eds.), *Learning and instruction: European research in an international context* (Vol. 2:2, pp. 295–308). Oxford, England: Pergamon.

Hieronymus, A. N., & Hoover, H. D. (1987). *Iowa tests of basic skills: Writing supplement teacher's guide*. Chicago: Riverside.

Hoetker, J. (1982). Essay examination topics and students' writing. *College Composition and Communication, 33*(4), 377-392.

Hoetker, J., & Brossell, G. (1989). The effects of systematic variations in essay topics on the writing performance of college freshmen. *College Composition and Communication, 40*(4), 414-421.

Hoover, M. R., & Politzer, R. L. (1981). Bias in composition tests with suggestions for a culturally appropriate assessment technique. In M. F. Whiteman (Ed.), *Writing: The nature, development, and teaching of written communication (pp. 197-208). Vol. 1: Variation in Writing: Functional and Linguistic-Cultural Differences*. Hillsdale, NJ: Lawrence Erlbaum Associates, Publishers.

Hudson, S. A., & Veal, L. R. (1981). *An empirical investigation of direct and indirect measure of writing*. (ERIC Document Reproduction Service No. ED205993)

Hughes, A. (1989). *Testing for Language Teachers*. Cambridge: Cambridge University Press.

Huot, B. (1990a). The literature of direct writing assessment: Major concerns and prevailing trends. *Review of Educational Research, 60*(2), 237-263.

Huot, B. (1990b). Reliability, validity and holistic scoring: What we know and what we need to know. *College Composition and Communication*, *41*(2), 201-213.

Jacobs, H., Zinkgraf, S., Wormuth, D., Hartfiel, V. F., & Hughey, J. (1981). *Testing ESL composition: A practical approach*. Rowley, MA: Newbury House.

Jenning, M., Fox, J., Graves, B., & Shohamy, E. (1999). The test-taker's choice: An investigation of the effect of topic on language test performance. *Language Testing, 16*(4), 426-456.

Jin, Y. (2001). *Xinli celiang* [Pengukuran psikologi]. Shanghai: Huadong Shifan Chubanshe.

Johnson, R. L., Penny, J., & Gordon, B. (2000). The relation between score resolution methods and interrater reliability: An empirical study of an analytic scoring rubric. *Applied Measurement in Education, 13*(2), 121-138.

Kamarudin Hj. Husin. (1988). *Pedagogi bahasa*. Petaling Jaya: Longman Malaysia Sdn. Bhd.

Kamarudin Hj. Husin. (1993). *Perkaedahan mengajar bahasa*. Kuala Lumpur: Kumpulan Budiman Sdn. Bhd.

Kamarudin Hj. Husin & Siti Hajar Hj. Abdul Aziz. (1997). *Penguasaan kemahiran menulis*. Subang Jaya: Kumpulan Budiman Sdn. Bhd.

Kane, M. T., & Brennan, R. L. (1977). *The generalizability of class means. Review of Educational Research, 47*(1), 267-292.

Kegley, P. H. (1986). The effect of mode of discourse on student writing performance: Implications for policy. *Educational Evaluation and Policy Analysis, 8*(2), 147-154.

Kellogg, R. T. (1987). Effects of topic knowledge on the allocation of processing time and cognitive effort to writing processes. *Memory & Cognition, 15*(3), 256-266.

Kellogg, R. T. (2007). Fundamentals of cognitive psychology. Los Angeles: Sage Publications, Inc.

Kementerian Pendidikan Malaysia. (1997). *Bahan pengajaran dan pembelajaran penulisan bahasa Cina KBSR tahap II*. Kuala Lumpur: Pusat Perkembangan Kurikulum.

Kementerian Pendidikan Malaysia. (1998). *Huraian Sukatan Pelajaran Kurikulum Bersepadu Sekolah Rendah: Bahasa Cina SJK(C) Tahun 4*. Kuala Lumpur: Pusat Perkembangan Kurikulum.

Kementerian Pendidikan Malaysia. (1999). *Huraian Sukatan Pelajaran Kurikulum Bersepadu Sekolah Rendah: Bahasa Cina SJK(C) Tahun 5-6*. Kuala Lumpur: Pusat Perkembangan Kurikulum.

Kementerian Pendidikan Malaysia. (2003a). *Pelaksanaan pengajaran dan pembelajaran Sains dan Matematik dalam Bahasa Inggeris di Sekolah Jenis Kebangsaan Cina SJK(C) mulai tahun 2003* (Surat Pekeliling Ikhtisas Bil. 12/2002). Kuala Lumpur: KPM.

Kementerian Pendidikan Malaysia. (2003b). *Sukatan Pelajaran Kurikulum Bersepadu Sekolah Rendah: Bahasa Cina*. Kuala Lumpur: Pusat Perkembangan Kurikulum.

Kementerian Pendidikan Malaysia. (2003c). *Sukatan Pelajaran Kurikulum Bersepadu Sekolah Rendah: Bahasa Cina* (versi bahasa Melayu, tidak diterbitkan). Kuala Lumpur: Pusat Perkembangan Kurikulum.

Khodori Hj. Ahmad & Jamil Adimin. (2003). *Memasyarakatan Pentaksiran*. Kertas kerja dibentangkan dalam Seminar Pentaksiran Pendidikan Kebangsaan 2003 anjuran Lembaga Peperiksan, Kementerian Pendidikan Malaysia (The Legend Hotel, Kuala Lumpur, 5-8 Mei, 2003).

Kinneavy, J. L. (1971). *A theory of discourse.* Englewood Cliffs, NJ: Prentice-Hall.

Klein, S.P., Stecher, B.M., Shavelson, R.J., McCaffrey, D., Ormseth, T., Bell, R. M., et al. (1998). Analytic versus holistic scoring of science performance tasks. *Applied Measurement in Education, 11*(2), 121-137.

Kobayashi, H., & Rinnert, C. (1996). Factors affecting evaluation in an EFL context: cultural rhetorical patterns and readers' background. *Language Learning, 46*(3), 397-437.

Koh, B. B. (1981). *Pengajaran Bahasa Malaysia*. Kuala Lumpur: Utusan Publications & Distributors Sdn. Bhd.

Koh, B. B. (1985). Analisis atas kesalahan ejaan dan tatabahasa Bahasa Malaysia yang dilakukan oleh sekumpulan siswazah. *Jurnal Dewan Bahasa, Ogos*. Kuala Lumpur: Dewan Bahasa dan Pustaka.

Koda, K. (1993). Task-induced variability in FL composition: language-specific perspectives. *Foreign Language Annals, 26*(3), 332-346.

Kondo-Brown, K. (2002). An analysis of rater bias with FACETS in measuring Japanese L2 writing performance. *Language Testing*, *19*(1), 1-29.

Kroll, B. (1998). Assessing writing abilities. *Annual Review of Applied Linguistics, 18*, 219-240.

Kubiszyn, T., & Borich, G. (2003). *Educational testing and measurement: Classroom application and practice*. New York: John Wiley & Sons, Inc.

Kunnan, A. J. (1992). An investigation of a criterion-referenced test using G-theory, and factor and cluster analyses. *Language Testing*, *9*(1), 30-49.

Lamb, H. (1987). *Student performance across the domain of school writing.* Paper presented at the Symposium on IEA Study of Written Composition at the Annual Meeting of the American Educational Research Association (Washington, DC, April 20-24, 1987). (ERIC Document Reproduction Service No. ED286194)

Lane, S., & Sabers, D. (1989). Using of generalizability theory for estimating the dependability of a scoring system for sample essays. *Applied Measurement in Education, 2*(3), 195-205.

Lee, Y-W., & Kantor, R. (2005). *Dependability of new ESL writing test scores: Evaluating prototype tasks and alternative rating schemes*. (ETS RR-05-14). Princeton, NJ: Educational Testing Service.

Lee, Y-W., Kantor, R., & Mollaun, P. (2002). *Score dependability of the writing and speaking section of New TOEFL.* Paper presented at the annual meeting of National Council on Measurement in Education (NCME), New Orleans, LA.

Lehmann, R. H. (1990). Reliability and generalizability of ratings of compositions. *Studies in Educational Evaluation, 16*, 501-512.

Lehmann, R. H. (1993). Rating the quality of student writing- findings from the IEA study of achievement in written composition. In H. Ari, S. Kari, & T. Sauli (Eds.), *Language testing: New openings* (pp. 186-204). Jyväskylä, Finland: Institute for Educational Research University of Jyväskylä.

Lembaga Peperiksaan Malaysia. (2002). *Manual pelaksanaan aktiviti pengesanan / pencerapan* (Tidak diterbitkan). Jalan Duta: Lembaga Peperiksaan Malaysia.

Lembaga Peperiksaan Malaysia. (2003). *Perealisasian item*. Jalan Duta: Lembaga Peperiksaan Malaysia. Tidak diterbitkan.

Lembaga Peperiksaan Malaysia. (2005). *Bahasa Cina UPSR: Format pentaksiran mulai 2005*. Shah Alam: Malindo Printers Sdn. Bhd.

Lembaga Peperiksaan Malaysia. (2008). *Pengumuman keputusan Ujian Pencapaian Sekolah Rendah (UPSR) Tahun 2008* (Tidak diterbitkan). Putrajaya: Lembaga Peperiksaan Malaysia.

Linacre, J. M. (1991). *Constructing measurement with a many-facet Rasch model*. Paper presented at the annual meeting of the American Educational Research Association (Chicago, IL, April 3-7, 1991). (ERIC Document Reproduction Service No. ED333047)

Linacre, J. M. (1999). Measurement of judgments. In G. N. Masters & J. P. Keeves (Eds.), *Advances in measurement in educational research and assessment* (pp. 244-253). Oxford: Elsevier Science Ltd.

Linn, R. L., Baker, E. L., & Dunbar, S. B. (1991). Complex, performance-based assessment. *Educational Researcher, 20*(8), 15-21.

Linn, R. L., & Gronlund, N. E. (2000). *Measurement and assessment in teaching* (8th ed.). New Jersey: Prentice-Hall, Inc.

Linn, R. L., & Gronlund, N. E. (2005). *Measurement and assessment in teaching* (9th ed.). New Jersey: Prentice-Hall, Inc.

Liu, Y. W., & Zhang, H. C. (1998). Application of generalizability theory in composition scoring. *Acta Psychological Sinica*, *30*(2), 211-218.

Lloyd-Jones, R. (1977). Primary trait scoring. In Cooper, C. R., & Odell, L. (Eds.), *Evaluating writing: Describing, measuring, judging* (pp. 32-66). Urbana, IL: National Council of teachers of English. (ERIC Document Reproduction Service No. ED143020)

Lumley, T., & McNamara, T. F. (1995). Rater characteristics and rater bias: Implications for training. *Language Testing*, *12*(1), 54–71.

Lunz, M. E., Wright, B. D., & Linacre, J. M. (1990). Measuring the impact of judge severity on examination scores. *Applied Measurement in Education*, *3*(4), 331–345.

Madaus, G. F., & O'Dwyer, L. M. (1999). A short history of performance assessment: Lessons learned. *Phi Delta Kappan, 80*(9), 688-695.

McAndrew, D. A. (1981). *The effect of an assigned rhetorical context on the holistic quality and syntax of the writing of high and low ability college writers*. (ERIC Document Reproduction Service No. ED235481)

McCann, T. M. (1989). Student argumentative writing: Knowledge and ability at three grade levels. *Research in the teaching of English, 23*(1), 62-76.

McNamara, T. (2000). *Language testing*. Oxford: Oxford University Press.

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

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50,* 741–749.

Milanovic, M., Saville, N., & Shuhong, S. (1993). A study of the decision-making behaviour of composition makers. In M. Milanovic & N. Saville (Eds.), *Performance testing, cognition and assessment: Selected papers from the 15th Language Testing Research Colloquium, Cambridge and Arnhem* (pp. 92-114). Cambridge: Cambridge University Press.

Miller, T. B., & Kane, M. (2001). The precision of change scores under absolute and relative interpretations. *Applied Measurement In Education, 14*(4), 307-327.

Miller, M. D., & Legg, S. M. (1993). Alternative assessment in a high-stakes environment. *Educational Measurement: Issues and Practice, 12*(2), 9-15.

Miller, D. M., & Linn, R. L. (2000). Validation of performance-based assessments. *Applied Psychological Measurement, 24*(4), 367-378.

Mohamad Sahari Nordin. (2002). *Pengujian dan penaksiran di bilik darjah*. Kuala Lumpur: Pusat Penyelidikan, Universiti Islam Antarabangsa Malaysia.

Mohd. Majid Konting. (1990). *Kaedah penyelidikan pendidikan*. Kuala Lumpur: Dewan Bahasa dan Pustaka.

Mohd. Najib Ghafar. (1997). *Pembinaan dan analisis ujian bilik darjah*. Sekudai: Penerbit Universiti Teknologi Malaysia.

Mokhtar Ismail. (1995). *Penilaian di bilik darjah*. Kuala Lumpur: Dewan Bahasa dan Pustaka.

Moon, T. R., & Hughes, K. R. (2002). Training and scoring issues involved in large-scale writing assessments. *Educational Measurement: Issues and Practice, 21*(2), 15-19.

Moss, P. A., Cole, N. S., & Khampalikit, C. (1982). A comparison of procedures to assess written language skills at grade 4, 7 and 10. *Journal of Educational Measurement, 19*(1), 37- 47.

National Assessment of Educational Progress. (1980). *Writing achievement, 1969-1979: Results from the third national writing assessment*. Denver, Colorado: NAEP.

Nadzri Isa. (2003). Kesan penggunaan bahan rangsangan dalam pengajaran karangan. *Dewan Bahasa, Jilid 3*(5)*, Mei.* Kuala Lumpur: Dewan Bahasa dan Pustaka.

Ng, S. N. (1991). *Pengukuran dan penilaian dalam pendidikan*. Kuala Lumpur: Penerbit Fajar Bakti Sdn. Bhd.

Nitko, A. J. (2004). *Educational assessment of students*. Upper Saddle River, NJ: Pearson Educational, Inc.

Norusis, M. J. (1999). *SPSS 9.0 guide to data analysis*. Upper Saddle River, NJ: Prentice-Hall, Inc.

Novak, J. R., Herman, J. L., & Gearhart, M. (1996). Establishing validity for performance-based assessments: An illustration for collections of student writing. *Journal of Educational Research, 89*(4), 220-233.

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd.). New York: McGraw- Hill, Inc.

Oller, J. W. (1979). *Language test at school*. London: Longman Group Ltd.

Pawlik, K. (2003). Theoretical perspective: Psychometrics. In Rocío Fernández-Ballesteros (Ed.), *Encyclopedia of psychological assessment* (Vol.2, pp. 1019-1023). London: Sage Publications Ltd.

Peterson, S. (2000). Fourth, sixth, and eighth graders' preferred writing topics and identification of gender markers in stories. *The Elementary School Journal, 101*(1), 79-100.

Pollitt, A., & Hutchinson, C. (1987). Calibrated graded assessments: Rasch partial credit analysis of performance in writing. *Language Testing, 4*(1), 72-92

Powers, D. E., & Fowles, M. E. (1999). Test-takers' judgments of essay prompts: Perceptions and performnace. *Educational Assessment, 6*(1), 3-22.

Prater, D. L. (1985). The effects of modes of discourse, sex of writer, and attitude toward task on writing performance in grade ten. *Educational and Psychological Research, 5*(4), 241-259.

Prater, D., & Padia, W. (1983). Effects of modes of discourse onwriting performance in grades four and six. *Research in the Teaching of English, 17*(2), 127-134.

Purves, A. (1992). Reflections on research and assessment in written composition. *Research in the Teaching of English, 26*(1), 109-123.

Purves, A. C., Soter, A., Takala, S., & Vahapassi, A. (1984). Towards a domain-referenced system for classifying composition assignments. *Research in the Teaching of English, 18*(4), 385-416.

Quek, W. K. (2003). *Mengajar untuk peperiksaan*. Kertas kerja dibentangkan dalam Seminar Pentaksiran Pendidikan Kebangsaan 2003 anjuran Lembaga Peperiksan, Kementerian Pendidikan Malaysia (The Legend Hotel, Kuala Lumpur, 5-8 Mei, 2003).

Quellmalz, E., Capell, F. J., & Chou, C. P. (1982). Effects of discourse and response mode on the measurement of writing competence. *Journal of Educational Measurement, 19*(4), 241-258.

Quellmalz, E. S. (1984). Toward successful large-scale writing assessment: Where are we now? Where do we go from here? *Educational Measurement: Issues and Practice, 3*(1), 29-35.

Quellmalz, E. S. (1990). Essay examination. In H. J. Walberg & G. D. Haertel (Eds.). *The international encyclopedia of educational evaluation* (pp. 510-515). Oxford, England: Pergamon Press.

Raimes, A. (1983). *Techniques in teaching writing*. Oxford: Oxford University Press.

Rajaratnam, N., Cronbach, L. J., & Gleser, G. C. (1965). Generalizability of stratified-parallel tests. *Psychometrika, 30*(1), 39-56.

Read, J. (1991). The validity of writing test tasks. In S. Anivan (Ed.), *Current development in language testing* (pp.77- 91). Singapore: SEAMEO Regional Language Centre.

Reed, W. M., Burton, J. K., & Kelly, P. P. (1985). The effects of writing ability and mode of discourse on cognitive capacity engagement. *Research in the Teaching of English, 19*(3), 283-297.

Reid, J. M. (1993). *Teaching ESL writing*. New Jersey: Prentice-Hall.

Roid, G. H. (1994). Patterns of writing skills derived from cluster analysis of direct-writing assessments. *Applied Measurement in Education, 7*(2), 159-170.

Romanoski, J., & Douglas, G. (2002). Test scores, measurement, and the use of analysis of variance: An historical overview. *Journal of Applied Measurement, 3*(3), 232-242.

Ruiz-primo, M. A., Baxter, G, P., & Shavelson, R. J. (1993). On the stability of performance assessments. *Journal of Educational Measurement, 30*(1), 41-53.

Ruth, L., & Murphy, S. (1984). Designing topics for writing assessment: Problems of meaning. *College Composition and Communication, 35*(4), 410-422.

Sachse, P. P. (1984). Writing assessment in Texas: Practices and problems. *Educational Measurement: Issues and Practice, 3*(1), 21-23.

Saddle, B., & Graham,S. (2005). The effects of peer-assisted sentence combining instruction on the writing performance of more and less skilled young writers. Journal of Educational Psychology, 97(1), 43-54.

Saville, N. (2003). The process of test development and revision within UCLES EFL. In C. Weir and M. Milanovic (Eds.), *Continuity and innovation: Revising the Cambridge Proficiency in English examination 1913-2002* (pp. 57-120). Cambridge: Cambridge University Press.

Schoonen, R. (2005). Generalizability of writing scores: an application of structural equation modeling. *Language Testing, 22*(1)*,* 1-30.

Schoonen, R., Vergeer, M., & Eiting, M. (1997). The assessment of writing ability: Expert readers versus lay readers. *Language Testing, 14*(2), 157-184.

Scott, V. M. (1996). *Rethinking foreign language writing*. Boston: Heinle & Heinle Publishers.

Shavelson, R. J., & Baxter, G. P., & Gao, X. (1993). Sampling variability of performance assessment, *Journal of Educational Measurement, 30*(3), 215-232.

Shavelson, R. J., & Webb, N. M. (1991). *Generalizability Theory: A primer.* Newbury Park, CA: Sage.

Shavelson, R. J., Webb, N. M., & Rowley, G. L. (1989). Generalizability Theory. *American Psychologist, 44*(6), 922–932.

Shi, L. (2001). Native and non-native speaking EFL teachers' evaluation of Chinese students' English writing. *Language Testing, 18*(3), 303-325.

Shi, L. (2003). Writing in two cultures: Chinese professors return from the West. *Canadian Modern Language Review, 59*(3), 369-391.

Shohamy, E., Gordon, C. M., & Kraemer, R. (1992) The effect of raters' background and training on the reliability of direct writing tests. *The Modern Language Journal*, *76*(1), 27-33.

Sidek Mohd. Noah. (2002). *Reka bentuk penyelidikan: Falsafah, teori dan praktis*. Serdang: Penerbi Universiti Putra Malaysia.

Smith, W. L., Hull, G. A., Land, R. E., Moore, M. T., Ball, C., Dunham, D. E., et al. (1985). Some effects of varying the structure of a topic on college students' writing. *Written Communication, 2*(1), 73-89.

Song, B., & Caruso, I. (1996). Do English and ESL faculty differ in evaluating the essays of native English-speaking and ESL students? *Journal of Second Language Writing, 5*(2), 163-182.

Spandel, V., & Stiggins, R. J. (1981). *Direct measures of writing skill: Issues and applications* (Revised ed.). Portland, Oregon: Northwest Regional Educational Laboratory. (ERIC Document Reproduction Service No. ED213035)

Spolsky, B. (1985). What does it mean to know how to use a language? An essay on the theoretical basis of language testing. *Language Testing, 2*(2), 180-191.

Spolsky, B. (1999). *Measured words*. Shanghai: Shanghai Foreign Language Educational Press.

Swartz, C. W., Hooper, S. R., Montgomery, J. W., Wakely, M. B., de Kruif, R. E. L., Reed, M., et al. (1999). Using Generalizability Theory to estimate the reliability of writing scores derived from holistic and analytical scoring methods. *Educational and Psychological Measurement, 59*(6), 492-506.

Swartz, R., Patience, W., & Whitney, D. R. (1985). *Adding an essay to the GED writing skills test: Reliability and validity issues* (GED Testing Service Research Studies No. 7). Washington, DC: American Council on Education. (ERIC Document Reproduction Service No. ED266288)

Sweigart, W. (1991). Classroom talk, knowledge development, and writing. *Research in the Teaching of English, 25*(4), 469-496.

Sweller, J. (1994). Cognitive load theory, learning difficulty, and instructional design. *Learning and Instruction, 4*(4), 295-312.

Stevens, J. J., & Clauser, P. (1996). *Longitudinal examination of a writing portfolio and the ITBS*. Paper presented at the Annual Meeting of the American Educational Research Association (New York, NY, April 8-12, 1996). (ERIC Document Reproduction Service No. ED397116)

Sultana, Q. (2001). *The university writing requirement: A study of the reliability of scores*. Paper presented at the annual meeting of the Mid-South Educational Research Association (30th, little Rock, AR, November 14-16, 2001). (ERIC Document Reproduction Service No. ED460147)

Suen, H. K. (1990). *Principles of test theories*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Thompson, B. (2003). Understanding reliability and coefficient alpha, really. In B. Thompson (Ed.), *Score reliability: Contemporary thinking on reliability issues* (pp. 3-23). London: Sage Publications.

Thomas, R. M. (2005). *High stakes testing: Coping with collateral damage*. Mahwah, NJ: Lawrence Erlbaum.

Thorkildsen, T. A. (2005). *Fundamentals of measurement in applied research*. Boston: Pearson Education, Inc.

Thorndike, R. M. (2005). *Measurement and evaluation in psychology and education* (7th ed.). Upper Saddle River, NJ: Pearson Education, Inc.

Tinsley, H. E. A., & Weiss, D. J. (2000). Interrater reliability and agreement. In H. E. A. Tinsley & S. D. Brown (Eds.), *Handbook of applied multivariate statistics and mathematical modeling* (pp. 95-124). San Diego, California: Academic Press.

Tobias, S. (1994). Interest, prior knowledge, and learning. *Review of Educational Research, 64(1)*, 37-54.

VanLeeuwen, D. M. (1997). Assessing reliability of measurements with generalizability theory: An application to inter-rater reliability. *Journal of Agricultural Education, 38(3)*, 36-42.

Wainer, H. (1993). Measurement problems. *Journal of Educational Measurement 30*(1), 1-21.

Wang, X. L. (2001). *Jiaoyu Celiang* [Pengukuran Pendidikan]. Shanghai: Universiti Perguruan Hua Dong.

Way, D. P., Joiner, E. G., & Seaman, M. A. (2000). Writing in the secondary foreign language classroom: The effects of prompts and tasks on novice learners of french. *The Modern Language Journal*, *84*(2), 171-184.

Webb, N. M., Schlackman, J., & Sugrue, B. (2000). The dependability and interchangeability of assessment methods in Science. *Applied Measurement in Education, 13*(3), 277-301.

Weigle, S. C. (1994). Effects of training on raters of ESL compositions. *Language Testing, 11*(2), 197-223.

Weigle, S. C. (1998). Using FACETS to model rater training effects. *Language Testing*, *15*(2), 263–287.

Weigle, S. C. (2002). *Assessing writing*. Cambridge: Cambridge University Press.

White, E. M. (1984). Holisticism. *College Composition and Communication, 35(4)*, 400-409.

White, E. M. (1994). *Teaching and assessing writing* (2nd ed.). San Francisco: Jossey-Bass.

Wiersma, W., & Jurs, S. G. (2005). *Research methods in education: An introduction* (8th ed.). Boston: Pearson Education, Inc.

Wigglesworth, G. (1994). Patterns or oral behavior in the assessment of an oral interaction test. *Australian Review of Applied Linguistics, 17(2),* 77-103.

Wiseman, S. (1949). The marking of English compositions in grammar school selection. *British Journal of Educational Psychology XIX 3*, 200-209.

Wolfe, E. W., Kao, C.-W., & Ranney, M. (1998). Cognitive differences in proficient and nonproficient essay scorers. *Written Communication*, *15*(4), 465-492.

Wood, R. (1991). *Assessment and testing: A survey of research*. Cambridge: Cambridge University Press.

Wu, C. F. Jeff., & Hamada. M. (2000). *Experiments: Planning, analysis, and parameter design optimization*. New York: John Wiley & Sons, Inc.

Yahya Othman. (2005). *Trend dalam pengajaran Bahasa Melayu*. Kuala Lumpur: PTS Professional Publishing Sdn. Bhd.

Yancey, K. B. (1999). Looking back as we look forward: Historicizing writing assessment. *College Composition & Communication, 50*(3), 483-503.

Yen, Fenliao. (2002). *Teachers College, Columbia University Working Papers in TESOL & Applied Linguistics, 4(1)*. Petik dari Clark, J., 1975. Theoretical and technical considerations in the oral test. *Language testing proficiency* (pp.10-

24). Retrieved December 20, 2007, from http://journals.tc-library.org/index.php/tesol/article/view/41/48

Yin, Y., & Shavelson, R. J. (2004). *Application of generalizability theory to concept-map assessment research.* (CSE Report 640). Los Angeles, CA: National Center for Research on Evaluation, Standards, and Student Testing (CRESST). CRESST. (ERIC Document Reproduction Service No. ED483407)

Zhu, X. H. (1990). Ertong zuowen nengli yinsu de yanjiu (Kajian tentang faktor-faktor kemahiran mengarang pelajar). *Journal of Hangzhou University*, *20*(3), 114-120.