

**PERBANDINGAN PENCAPAIAN MURID DALAM UJIAN SAINS
BERDASARKAN VERSI BAHASA INGGERIS
DAN BAHASA MELAYU**

**(COMPARING STUDENTS' ACHIEVEMENT FOR SCIENCE TEST
IN ENGLISH AND IN MALAY LANGUAGE)**

Lim Peh Hwah

SMJK Heng Ee, Jalan Hamilton,
11600 Georgetown, Pulau Pinang
E-mel: phlim167@yahoo.com

Ong Saw Lan

Pusat Pengajian Ilmu Pendidikan
Universiti Sains Malaysia, 11800 USM Pulau Pinang
E-mel: osl@usm.my

Abstrak: Kementerian Pendidikan Malaysia telah memperkenalkan dasar mengajar Sains dan Matematik dalam bahasa Inggeris. Hasil daripada perubahan dasar bahasa, Lembaga Peperiksaan Malaysia dan Majlis Peperiksaan Malaysia menggunakan ujian dwibahasa (iaitu bahasa Inggeris dan bahasa Melayu) untuk mentaksirkan kedua-dua mata pelajaran tersebut pada semua peringkat peperiksaan awam. Kajian telah menunjukkan ujian dalam bahasa yang berlainan mungkin tidak menunjukkan kesetaraan psikometrik. Dengan itu, kajian ini dijalankan untuk menentukan punca perbezaan pencapaian disebabkan oleh perbezaan dalam bahasa atau perbezaan kebolehan pelajar yang sebenarnya.

Kata kunci: bahasa pentaksiran, kesetaraan ujian, kesetaraan pencapaian

Abstract: The Malaysian Ministry of Education implemented the policy of teaching Science and Mathematics in English. As a result of this change in language policy, the Malaysia Examination Board and Malaysia Examination Council use bilingual test (English and Malay language) to assess the two subjects at all levels of the public examinations. Studies have shown that tests in different languages may not exhibit psychometric equivalence. This study is conducted to determine whether sources of differential achievement is a result of language differences or real difference due to students' ability.

Keywords: language of assessment, test equivalence, performance equivalence

LATAR BELAKANG

Kementerian Pendidikan Malaysia telah memperkenalkan pengajaran Matematik dan Sains dalam bahasa Inggeris pada tahun 2003. Semua pelajar sekolah rendah

dan sekolah pendidikan khas tahun satu wajib mempelajari Matematik dan Sains dalam bahasa Inggeris. Pelajar di Sekolah Rendah Kebangsaan (SRK) mempelajari Sains dan Matematik dalam bahasa Inggeris manakala pelajar di Sekolah Jenis Kebangsaan Cina (SJKC) menggunakan kedua-dua bahasa Cina dan bahasa Inggeris. Di sekolah menengah pula, pelajar Tingkatan Satu mempelajari Matematik dan Sains dalam bahasa Inggeris. Sementara pelajar Tingkatan Enam Bawah mempelajari Matematik, Biologi, Kimia, Fizik serta perkomputeran dalam bahasa Inggeris. Dengan perubahan dasar bahasa ini, Lembaga Peperiksaan Malaysia (LPM) dan Majlis Peperiksaan Malaysia memperkenalkan ujian dwibahasa (iaitu bahasa Inggeris dan bahasa Melayu) untuk mentaksirkan mata pelajaran tersebut pada semua peringkat peperiksaan awam.

Rasional LPM menyediakan ujian dwibahasa dalam tempoh transisi (dari tahun 2003 hingga tahun 2007) ialah supaya pelajar dapat memahami kehendak soalan dengan baik. Murid diberi pilihan untuk menjawab sama ada dalam bahasa Inggeris atau bahasa Melayu, atau kedua-duanya. Menurut Ketua Pengarah Pelajaran, Datuk Ahmad Sipon, penggunaan dwibahasa dalam soalan mata pelajaran Matematik dan Sains diadakan sehingga 2007. Mulai tahun 2008, ujian kebangsaan Matematik dan Sains dicadangkan hanya disediakan dalam bahasa Inggeris (The Star, 2005).

MASALAH

Penggunaan bahasa Inggeris sebagai bahasa pengantar untuk mata pelajaran Sains dan Matematik telah membimbangkan para pendidik, ibu bapa dan pelajar. Pelajar terpaksa belajar konsep-konsep Sains dan Matematik dalam bahasa yang belum dikuasai. Tambahan pula, guru Sains dan Matematik dalam perkhidmatan telah dilatih mengajar mata pelajaran tersebut dalam bahasa Melayu. Sebahagian ibu bapa dan pendidik tidak menyambut baik penggunaan bahasa Inggeris sebagai bahasa pengantar Matematik dan Sains kerana mereka berpendapat bahawa pembelajaran melalui bahasa ibunda adalah lebih berkesan.

Selain daripada mata pelajaran Sains dan Matematik, mata pelajaran lain seperti Sejarah, Geografi dan Kemahiran Hidup masih tetap diajar dalam bahasa Melayu. Kementerian Pendidikan Malaysia akan membuat keputusan untuk tidak menggunakan dwibahasa dalam peperiksaan umum bagi Sains dan Matematik pada penghujung tahun 2008. Apabila Sains dan Matematik ditaksir dalam bahasa Inggeris, isu kesahan perlu ditangani. Dengan itu, untuk mengenalpasti sama ada pencapaian pelajar dalam mata pelajaran Sains dipengaruhi oleh faktor bahasa, ujian sains disediakan dalam dua bahasa yang berlainan tetapi isi

kandungan yang sama ditadbirkan kepada calon untuk membolehkan perbandingan dilakukan.

Perbandingan pencapaian pelajar berdasarkan ujian sains yang ditadbir dalam bahasa yang berlainan mungkin tidak memberi maklumat yang jitu serta tidak adil. Kajian telah menunjukkan ujian dalam bahasa yang berlainan mungkin tidak menunjukkan kesetaraan psikometrik (Angoff & Cook, 1988; Geisinger, 1994; Hambleton, 1994; Olmedo, 1981; Prieto, 1992). Oleh itu, adalah penting untuk menentukan kesetaraan dua ujian yang berlainan bahasa sebelum perbandingan pencapaian pelajar dibuat. Tanpa mengenalpasti kesetaraan ujian dalam bahasa yang berlainan, maka adalah tidak sah untuk menentukan punca kepada perbezaan pencapaian, sama ada disebabkan oleh perbezaan dalam bahasa atau kebolehan pelajar yang sebenarnya.

TEORI PENYETARAAN UJIAN

Untuk menganalisis item ujian, dua jenis item statistik yang biasa telah digunakan (Zhu, 1998). Jenis pertama dikenali sebagai item statistik klasik, iaitu kesukaran item dan indeks diskriminasi. Namun, indeks jenis ini mempunyai kelemahan iaitu nilai-nilainya bergantung kepada populasi. Keputusan analisis berubah apabila kumpulan kajian adalah berlainan disebabkan pengetahuan dan tahap kemahiran sampel yang berlainan (Woldbeck, 1998).

Jenis kedua item statistik dihasilkan daripada penentukuran model Respon Item yang mengandungi statistik kesukaran item, statistik tentukur ralat dan statistik kesepadanan item (Kolen & Brennan, 1995). Selain daripada itu, mereka juga menyatakan bahawa statistik kesepadanan item memberi anggaran sejauh mana sesuatu item mematuhi jangkaan model, iaitu calon yang berpengetahuan mempunyai kebarangkalian yang lebih tinggi untuk memberi jawapan yang betul. Penyetaraan ujian berdasarkan Teori Respon Item adalah berdasarkan perbandingan parameter yang dianggar secara berasingan kemudian diletakkan pada satu skala sepunya (Woldbeck, 1998).

Analisis Keberbezaan Fungsi Item (DIF) biasanya digunakan semasa proses pengadaptasian untuk mengenalpasti item yang mempunyai fungsi berlainan di antara kumpulan yang menggunakan bahasa berlainan. (Allalouf, Hambleton & Sireci, 1999; Budgell, Raju & Quartetti, 1995; Hambleton, 1994; van de Vijver & Leung, 1997). DIF dikatakan wujud apabila sampel dari kumpulan bahasa yang berlainan mempunyai kebarangkalian yang berlainan untuk menjawab sesuatu item dengan betul selepas mengambil kira kebolehan keseluruhan.

Menurut Abedi (2006) pembelajaran dan pentaksiran *English Language Learner* dengan menggunakan bahasa Inggeris telah menyekat pencapaian pelajar dalam matematik. Dengan itu, kecekapan bahasa menjejaskan kebolehpercayaan dan kesahan sesuatu pengujian. Stansfield (1996) mencadangkan penggunaan bahasa ibunda sebagai bahasa pengujian untuk mengatasi masalah tersebut. Namun, Abedi (2006) berpendapat bahawa menterjemahkan item ujian daripada bahasa Inggeris kepada bahasa ibunda tidak memberi manfaat kepada pelajar sekiranya bahasa pengantar adalah bahasa Inggeris. Lord (1980) juga menyuarakan pandangan yang sama dengan menyatakan bahawa item dalam bahasa ibunda telah mengelirukan pelajar yang mempelajari konsep dalam bahasa Inggeris kerana mereka mungkin tidak biasa dengan terminologi yang digunakan.

Dengan itu, kajian ini meneliti kesan pengujian sains dalam bahasa Melayu, iaitu bahasa pengantar utama, terhadap pencapaian sains. Kajian ini terdiri daripada dua Fasa: fasa pertama meneliti kesetaraan dua ujian sains dalam bahasa yang berlainan. Kesetaraan ditentukan dengan indeks diskriminasi, indeks kesukaran, kebolehpercayaan serta DIF item. Lazimnya, dua ujian yang mengukur kandungan yang sama, juga mengukur konstruk yang sama. Tetapi, dalam kajian ini, penggunaan bahasa yang berlainan mungkin menyebabkan konstruk yang berlainan diukur. Fasa kedua kemudian membandingkan pencapaian pelajar yang ditadbirkan ujian sains dalam bahasa yang berlainan.

RASIONAL

Pada umumnya, masyarakat di Malaysia amat mementingkan pencapaian peperiksaan terutamanya peperiksaan awam. Hal ini adalah disebabkan peperiksaan tersebut memainkan peranan yang penting dalam sistem pentaksiran di Malaysia. Selain daripada peperiksaan awam, ujian dan peperiksaan di peringkat sekolah seperti ujian penggal juga amat diberi penekanan oleh pihak sekolah serta ibu bapa. Walau bagaimanapun, ujian mata pelajaran sains pada peringkat sekolah dikendalikan dalam bahasa Inggeris sahaja dan bukan dwibahasa seperti dalam peperiksaan awam. Dengan itu, pencapaian pelajar pada peringkat sekolah mungkin tidak sejajar dengan pencapaian peperiksaan awam. Walaupun mata pelajaran Sains diajar dalam bahasa Inggeris, pelajar yang lemah dalam bahasa Inggeris cenderung untuk memberi jawapan dalam bahasa Melayu. Pelajar mungkin tidak dapat menunjukkan pencapaian sains yang sebenar jika menjawab dalam bahasa yang belum dikuasai dengan sepenuhnya.

TUJUAN DAN OBJEKTIF

Kajian ini bertujuan menentukan kesan bahasa terhadap pencapaian pelajar dalam mata pelajaran Sains dengan menghubungkan dua ujian yang berlainan bahasa (bahasa Inggeris dan bahasa Melayu) pada satu skala pengukuran yang seponya. Secara lebih spesifik, tujuan utama kajian ini adalah untuk mengkaji sama ada wujud perbezaan pencapaian antara pelajar yang menjawab item-item Sains dalam bahasa Inggeris dengan pelajar yang menjawab item Sains dengan menggunakan bahasa Melayu. Secara khususnya objektif kajian ini adalah:

- (1) Mengkaji kesetaraan ujian dalam bahasa berlainan dari segi kebolehan diskriminasi, aras kesukaran, kebolehppercayaan, serta DIF.
- (2) Mengkaji sama ada wujudnya perbezaan pencapaian Sains antara pelajar yang menjawab ujian dalam versi bahasa Inggeris dengan versi bahasa Melayu.

PERSOALAN

Beberapa soalan perlu dijawab untuk mencapai tujuan kajian ini.

- (1) Sejauh manakah item versi bahasa Inggeris setara dengan item versi bahasa Melayu dari segi
 - (a) kebolehan diskriminasi?
 - (b) aras kesukaran?
 - (c) kebolehppercayaan?
- (2) Adakah wujud DIF di antara kedua-dua versi ujian?
- (3) Adakah terdapat perbezaan yang signifikan di antara pencapaian pelajar yang menjawab ujian versi bahasa berlainan?

METHODOLOGI

Reka Bentuk Kajian

Kajian ini menggunakan reka bentuk kumpulan rambang setara. Setiap kumpulan ditadbir instrumen dalam bahasa yang berlainan. Dua ujian sains yang berlainan

dari segi bahasa iaitu bahasa Inggeris dan bahasa Melayu ditadbirkan kepada kedua-dua kumpulan pelajar sekali sahaja.

Untuk mendapatkan dua kumpulan yang setara, pelajar-pelajar dalam setiap kelas telah dibahagikan kepada dua kumpulan mengikut nombor yang diberi. Pelajar yang diberi nombor ganjil menjawab ujian versi bahasa Inggeris, sementara pelajar yang bernombor genap pula menjawab ujian versi bahasa Melayu.

Instrumen Kajian

Kajian ini menggunakan ujian Sains dalam dua versi bahasa yang berlainan, iaitu bahasa Inggeris dan bahasa Melayu. Kedua-dua versi ujian itu mengandungi item-item objektif berbentuk aneka-pilihan. Item-item dalam ujian versi bahasa Inggeris telah dipilih daripada item-item TIMSS (Third International Mathematics and Sciences Study) yang dikeluarkan pada tahun 1999 dan tahun 2003, dan bersesuaian dengan sukatan pelajaran Sains Tingkatan Satu dan Tingkatan Dua di Malaysia.

Terdapat sejumlah 146 item dikeluarkan oleh TIMSS untuk tahun 1999 dan tahun 2003. Seorang guru Sains telah membantu mengenalpasti 40 item yang sepadan dengan sukatan pelajaran Sains Tingkatan Satu dan Tingkatan Dua. Dua orang guru Sains yang berpengalaman mengajar mata pelajaran Sains selama 15 tahun dan juga merupakan pemeriksa Sains untuk Penilaian Menengah Rendah (PMR) kemudian melengkapkan Jadual Penentuan Ujian untuk item-item yang dipilih. Mereka memadankan item dengan sukatan pelajaran Sains Tingkatan Satu atau Tingkatan Dua. Item-item yang didapati tidak sepadan dengan sukatan pelajaran dikeluarkan. Kedua-dua orang guru Sains mengelompokkan item-item kepada sebelas tajuk, lima untuk Tingkatan Satu dan enam untuk Tingkatan Dua. Tajuk-tajuk Tingkatan Satu termasuk; (i) Sel Sebagai Unit Asas Kehidupan, (ii) Jirim, (iii) Kepelbagaian Sumber di Bumi, (iv) Udara di Sekeliling Kita, dan (v) Sumber Tenaga. Manakala tajuk Tingkatan Dua adalah (i) Haba, (ii) Dunia Melalui Deria Kita, (iii) Nutrisi, (iv) Biokepelbagaian, (v) Air dan Larutan, dan (vi) Mesin Ringkas. Berdasarkan Jadual Spesifikasi Ujian yang dibina, mereka mengenalpasti 10 item adalah kurang sesuai dengan tajuk Sains Tingkatan Satu dan Tingkatan Dua. Ujian akhir Sains terdiri daripada 30 item sahaja.

Untuk menentukan kesahan isi kandungan ujian, tiga orang guru Sains diminta menilai kesesuaian item untuk menguji topik-topik sukatan pelajaran Sains menggunakan skala *Likert*, iaitu daripada sangat sesuai (5) kepada tidak sesuai (1) untuk menentukan kesesuaian item mengukur tajuk yang dipadankan. Sekiranya min sesuatu item kurang daripada 2.5, maka item itu adalah tidak sesuai untuk menguji tajuk-tajuk berkenaan. Hasil penilaian ketiga-tiga orang

guru sains menunjukkan bahawa kesemua item adalah sesuai untuk tajuk yang disenaraikan.

Setelah mengenal pasti item untuk versi bahasa Inggeris, item tersebut kemudian diterjemah dan diadaptasi kepada versi bahasa Melayu. Penterjemahan dilakukan serentak oleh dua orang guru sains dan seorang guru bahasa yang fasih dalam bahasa Inggeris dan bahasa Melayu. Ketiga-tiga penterjemah kemudian mengadakan perbincangan dan bersetuju dengan ujian versi bahasa Melayu yang diterjemah.

Persampelan

Sepuluh buah sekolah di sekitar Pulau Pinang terlibat dalam kajian; lima adalah Sekolah Menengah Kebangsaan dan lima lagi Sekolah Menengah Jenis Kebangsaan. Sekolah-sekolah yang terpilih telah menjalankan peperiksaan akhir tahun sekolah dengan merangkumi sukatan pelajar Tingkatan Satu dan Tingkatan Dua. Mengikut kriteria TIMSS gred 8, pelajar yang terlibat mesti melalui sekurang-kurangnya lapan tahun pendidikan formal. Oleh itu, sampel kajian ini terdiri daripada pelajar Tingkatan Dua sahaja.

Jumlah sampel dalam kajian ini adalah 1000 orang pelajar. Setiap pelajar hanya menjawab satu versi ujian sains. Seramai 500 orang pelajar telah menjawab item dalam versi bahasa Inggeris manakala 500 orang pelajar lagi menjawab item dalam versi bahasa Melayu.

Sampel yang dipilih untuk kajian ini terdiri daripada pelajar yang mempunyai kecekapan minimum membaca dan memahami bahasa Inggeris dan bahasa Melayu berdasarkan klasifikasi oleh guru kelas.

Prosedur Pentadbiran Ujian

Kajian ini dijalankan seminggu selepas peperiksaan akhir tahun iaitu pada akhir bulan Oktober dan awal bulan November 2007. Dengan itu, pelajar dianggap telah membuat ulang kaji untuk tajuk sains yang berkaitan. Semasa Malaysia mengambil bahagian dalam ujian TIMSS gred 8 tahun 2003, ujian juga ditadbir pada akhir bulan Oktober dan awal bulan November 2003.

Ujian ini dijalankan pada hari yang berlainan untuk sekolah yang berlainan mengikut aktiviti dan kebenaran sekolah. Tempoh seminggu digunakan untuk mentadbir ujian di kesemua sekolah.

Walaupun dalam ujian Penilaian Menengah Rendah (PMR), terdapat sejumlah 40 item dan masa yang diperuntukkan adalah satu jam. Agihan masa ujian adalah

berdasarkan jumlah item dalam setiap versi ujian. Jumlah masa untuk menjawab setiap versi ujian adalah 45 minit kerana ujian mempunyai 30 item sahaja.

Versi ujian yang berlainan iaitu ujian versi bahasa Inggeris dan versi bahasa Melayu diberikan kepada pelajar mengikut kedudukan nombor dalam kelas. Pelajar yang bernombor ganjil menjawab ujian versi bahasa Inggeris dan nombor genap pula menjawab ujian versi bahasa Melayu. Dengan itu penyelidik dapat mengumpul data daripada dua kumpulan yang setara secara rambang.

Semasa pentadbiran ujian, penyelidik dan guru-guru sains di sekolah berkenaan membaca arahan tentang bilangan soalan terkandung dalam ujian dan jangka masa yang diperuntukkan. Pelajar menjawab dalam kertas jawapan yang disediakan bersamaan dengan kertas soalan. Dengan adanya penerangan daripada guru dan penyelidik, masalah seperti pelajar tidak sempat menjawab kesemua soalan serta kecuaiian seperti tidak menjawab soalan pada muka surat akhir dapat dielakkan. Sebaik sahaja tempoh masa tamat, semua kertas soalan dan kertas jawapan dikutip balik.

ANALISIS DATA

Data dianalisis untuk memperoleh nilai indeks diskriminasi, indeks kesukaran dan seterusnya kebolehpercayaan bagi kedua-dua ujian menggunakan perisian komputer TAP (Test Analysis Program). Keberbezaan fungsi item dikenalpasti menggunakan regresi logistik dan Model Rasch. Perbandingan di antara pencapaian pelajar yang menjawab ujian dalam bahasa Inggeris dan bahasa Melayu dilakukan menggunakan statistik deskriptif serta lengkung taburan.

Perbandingan Kesetaraan Item

Indeks diskriminasi

Untuk menentukan sama ada sesuatu item dapat berfungsi dengan baik dalam sesuatu ujian, nilai indeks diskriminasi telah ditentukan. Nilai indeks 0.40 atau lebih merupakan item yang sangat baik, 0.30 hingga 0.39 merupakan item yang baik tetapi boleh diperbaiki lagi, 0.20 hingga 0.29 merupakan item yang sederhana dan memerlukan pengubahsuaian, item yang nilai indeks diskriminasinya berada di bawah 0.19 merupakan item yang lemah dan perlu banyak pengubahsuaian atau dikeluarkan (Ebel & Frisbie, 1986).

Indeks kesukaran

Kesukaran item merupakan peratus pelajar yang mengambil ujian menjawab sesuatu item itu dengan tepat. Semakin tinggi peratus sesuatu item dijawab dengan betul oleh pelajar, bermaksud semakin senang sesuatu item. Semakin tinggi indeks kesukaran, semakin senang item itu (Wood, 1960). Untuk mendapatkan kesukaran item, bahagikan bilangan pelajar yang menjawab betul untuk sesuatu item dengan jumlah pelajar yang menjawab item tersebut. Nisbah untuk item itu biasanya dilabel sebagai p dan dipanggil sebagai kesukaran item (Crocker & Algina, 1986).

Indeks kebolehpercayaan

Dalam kajian ini, KR-20 (Kuder-Richardson 20) telah digunakan untuk menentukan kebolehpercayaan item kerana KR-20 sesuai digunakan untuk ujian aneka pilihan. KR-20 digunakan untuk mengukur kebolehpercayaan ketekalan dalaman atau sebaik mana ujian itu mengukur satu faktor kognitif. Indeks adalah dalam lingkungan 0.00 hingga 1.00. Nilai yang hampir kepada 0.00 menunjukkan terdapat banyak faktor yang tidak diketahui dan bukan faktor yang hendak diukur. Sebaliknya, jika nilai hampir kepada 1.00, maka bermaksud terdapat satu faktor yang diukur.

Keberbezaan Fungsi Item (DIF)

Sebelum DIF dikenalpasti, keserasian data dengan Model Rasch ditentukan terlebih dahulu. Dalam kajian ini, nilai *INFIT* dan *OUTFIT* min kuasa dua dalam julat antara 0.70–1.30 digunakan sebagai julat yang sesuai seperti yang dicadangkan oleh Bond dan Fox (2001). Walau bagaimanapun dalam kajian ini, keutamaan diberi kepada isi kandungan item yang sesuai untuk membuat keputusan sama ada mengeluarkan item atau tidak.

Setelah mengenalpasti nilai *INFIT* dan *OUTFIT* min kuasa dua untuk kedua-dua versi item ujian, DIF dapat ditentukan. Linacre (2005) menyatakan bahawa item-item yang $t \geq 1.96$ merupakan item yang menunjukkan keberbezaan fungsi item pada aras $p < 0.05$.

Perbandingan Pencapaian Pelajar

Perbandingan pencapaian pelajar dalam ujian yang berlainan bahasa iaitu bahasa Inggeris dan bahasa Melayu dilakukan dengan melihat kepada taburan skor. Ini ditunjukkan dengan menggunakan graf bar, penggunaan z-skor untuk graf garisan, serta analisis statistik deskriptif.

Dapatan Kajian

Analisis untuk indeks diskriminasi ditunjukkan dalam Jadual 1. Dua item (item 6 dan item 9) untuk versi bahasa Melayu dan satu item versi bahasa Inggeris (item 9) mempunyai indeks diskriminasi kurang daripada 0.20. Oleh itu, secara keseluruhan indeks diskriminasi untuk kedua-dua versi ujian mempunyai indeks diskriminasi yang memuaskan dan baik berdasarkan kriteria yang dikemukakan oleh Ebel dan Frisbie (1986), iaitu nilai indeks diskriminasi sama atau melebihi 0.20.

Jadual 1. Perbandingan indeks diskriminasi

Item	Ujian Sains dalam bahasa Melayu	Ujian Sains dalam bahasa Inggeris	Perbezaan
1	0.44	0.41	0.03
2	0.31	0.49	-0.18
3	0.34	0.43	-0.09
4	0.56	0.58	-0.02
5	0.51	0.34	0.17
6	0.15*	0.33	-0.18
7	0.38	0.26	0.12
8	0.38	0.55	-0.17
9	0.09*	0.04*	0.05
10	0.58	0.49	0.09
11	0.34	0.42	-0.08
12	0.61	0.60	0.01
13	0.28	0.37	-0.09
14	0.28	0.33	-0.05
15	0.66	0.63	0.03
16	0.44	0.42	0.02
17	0.55	0.59	-0.04
18	0.46	0.51	-0.05
19	0.62	0.59	0.03
20	0.31	0.48	-0.17
21	0.47	0.35	0.12
22	0.29	0.44	-0.15
23	0.41	0.56	-0.15
24	0.65	0.59	0.06

(bersambung)

Jadual 1. (sambungan)

Item	Ujian Sains dalam bahasa Melayu	Ujian Sains dalam bahasa Inggeris	Perbezaan
25	0.61	0.63	-0.02
26	0.48	0.66	-0.18
27	0.22	0.20	0.02
28	0.66	0.62	0.04
29	0.38	0.43	-0.05
30	0.58	0.57	0.01

* Indeks diskriminasi yang kurang daripada 0.20.

- Nilai negatif wujud apabila indeks diskriminasi untuk item bahasa Melayu lebih rendah daripada indeks diskriminasi untuk item bahasa Inggeris.

Hasil analisis indeks kesukaran item ditunjukkan dalam Jadual 2.

Jadual 2. Perbandingan Aras Kesukaran Item

Item	Bahasa Melayu	Bahasa Inggeris	Perbezaan
1	0.75	0.70	0.05
2	0.64	0.58	0.06
3	0.84	0.81	0.03
4	0.65	0.61	0.04
5	0.53	0.49	0.04
6	0.92	0.88	0.04
7	0.74	0.71	0.03
8	0.55	0.48	0.07
9*	0.18*	0.18*	0
10	0.53	0.50	0.03
11	0.87	0.83	0.04
12	0.55	0.51	0.04
13*	0.73	0.73	0
14	0.82	0.76	0.06
15	0.57	0.58	-0.01
16	0.46	0.45	0.01
17	0.67	0.63	0.04
18	0.42	0.40	0.02
19	0.35	0.36	-0.01
20	0.73	0.66	0.07

(bersambung)

Jadual 2. (sambungan)

Item	Bahasa Melayu	Bahasa Inggeris	Perbezaan
21	0.54	0.49	0.05
22*	0.54	0.54	0
23	0.65	0.60	0.05
24	0.70	0.72	-0.02
25*	0.61	0.61	0
26	0.79	0.65	0.14
27	0.50	0.48	0.02
28*	0.68	0.68	0
29	0.86	0.84	0.02
30	0.40	0.35	0.05

* Item yang mempunyai aras kesukaran yang sama untuk kedua-dua versi ujian.

– Nilai negatif wujud apabila aras kesukaran untuk item bahasa Melayu lebih rendah daripada aras kesukaran untuk item bahasa Inggeris.

Perbezaan indeks diskriminasi item antara kedua-dua versi juga ditunjukkan dalam Jadual 1. Didapati perbezaannya adalah tidak ketara. Perbezaan yang paling besar ialah 0.18, iaitu item 2, item 6 dan item 26. Terdapat 10 item yang menunjukkan perbezaan indeks diskriminasi yang melebihi 0.10, iaitu item 2, item 5, item 6, item 7, item 8, item 20, item 21, item 22, item 23 dan item 26. Tujuh belas item mempunyai perbezaan indeks diskriminasi kurang daripada 0.10.

Item 9 mempunyai indeks kesukaran yang paling rendah untuk versi bahasa Melayu dan versi bahasa Inggeris, iaitu nilai 0.18. Item ini merupakan item yang paling sukar kerana hanya 18% pelajar menjawab item ini dengan betul. Sebaliknya, item yang paling senang ialah item 6 untuk versi bahasa Melayu dan bahasa Inggeris, iaitu indeks kesukaran adalah 0.92 untuk versi bahasa Melayu dan 0.88 untuk versi bahasa Inggeris.

Sembilan item untuk versi bahasa Melayu mempunyai nilai indeks kesukaran antara 0.7–0.9, iaitu item 1, item 3, item 7, item 11, item 13, item 14, item 20, item 26 dan item 29. Untuk versi bahasa Inggeris, 8 item mempunyai nilai indeks kesukaran dalam julat 0.7–0.9. Item-item ini adalah item 3, item 6, item 7, item 11, item 13, item 14, item 24 dan item 29. Item-item dengan nilai indeks kesukaran ini dikelaskan sebagai item yang senang.

Kebanyakan item, sejumlah 19 item versi bahasa Melayu dan 21 item untuk versi bahasa Inggeris adalah dalam aras kesukaran yang memuaskan, iaitu antara 0.2–0.7.

Perbandingan nilai indeks kesukaran item bahasa Melayu dan bahasa Inggeris pada turus ketiga dalam Jadual 2, menunjukkan perbezaan indeks kesukaran item di antara kedua-dua versi adalah kurang daripada 0.1 kecuali item 26 dengan perbezaan 0.14. Tambahan lagi, lima item mempunyai indeks kesukaran yang sama untuk kedua-dua versi bahasa, iaitu item 9, item 13, item 22, item 25 dan item 28. Dua item versi bahasa Melayu mempunyai indeks kesukaran yang lebih kecil daripada item dalam versi bahasa Inggeris, iaitu item 19 dan item 24. Item-item ini seolah-olah lebih susah dalam bahasa Melayu. Namun, 23 item versi bahasa Melayu mempunyai nilai indeks kesukaran yang lebih tinggi daripada item dalam versi bahasa Inggeris. Ini menunjukkan bahawa kebanyakan item adalah lebih mudah dalam bahasa Melayu.

Kebolehpercayaan keseluruhan untuk ujian versi bahasa Melayu ialah 0.82 dan 0.85 untuk versi bahasa Inggeris.

Jadual 3. Perbandingan kebolehpercayaan

Item	KR-20 bahasa Melayu jika item dikeluarkan	KR-20 bahasa Inggeris jika item dikeluarkan
1	0.814	0.841
2	0.82	0.842
3	0.815	0.838
4	0.811	0.838
5	0.815	0.844
6	0.819	0.84
7	0.817	0.845+
8	0.818	0.84
9	0.824+	0.850+
10	0.813	0.841
11	0.814	0.838
12	0.812	0.838
13	0.82	0.843
14	0.818	0.842
15	0.81	0.837
16	0.817	0.843
17	0.812	0.837
18	0.816	0.841
19	0.811	0.838

(bersambung)

Jadual 3. (sambungan)

Item	KR-20 bahasa Melayu jika item dikeluarkan	KR-20 bahasa Inggeris jika item dikeluarkan
20	0.818	0.841
21	0.816	0.844
22	0.822+	0.843
23	0.816	0.84
24	0.808	0.836
25	0.811	0.836
26	0.811	0.835
27	0.824+	0.849+
28	0.808	0.836
29	0.813	0.838
30	0.813	0.839

+ Menunjukkan (KR-20) bertambah jika item tersebut dikeluarkan.

- Nilai negatif wujud apabila aras kesukaran untuk item bahasa Melayu lebih rendah daripada aras kesukaran untuk item bahasa Inggeris.

Jadual 4. Keberbezaan fungsi item (DIF) menggunakan perisian WINSTEPS

No. Item	Pengukuran DIF	DIF S.E.	Pengukuran DIF	DIF S.E.	DIF Contrast	Joint S.E.	t	d.f
b. Inggeris		b. Melayu		Perbezaan				
1	-0.54	0.11	-0.65	0.11	0.11	0.16	0.72	995
2	0.11	0.1	-0.02	0.1	0.14	0.14	0.96	995
3	-1.23	0.12	-1.32	0.13	0.09	0.18	0.52	995
4	-0.04	0.1	-0.07	0.1	0.02	0.15	0.17	995
5	0.58	0.1	0.56	0.1	0.02	0.14	0.13	995
6	-1.87	0.15	-2.13	0.17	0.26	0.22	1.16	995
7	-0.62	0.11	-0.58	0.11	-0.04	0.16	-0.28	994
8	0.61	0.1	0.45	0.1	0.16	0.14	1.12	995
9	2.42	0.13	2.54	0.13	-0.12	0.18	-0.69	994
10	0.53	0.1	0.54	0.1	-0.01	0.14	-0.09	994
11	-1.4	0.13	-1.54	0.14	0.14	0.19	0.74	995
12	0.47	0.1	0.44	0.1	0.03	0.14	0.19	995

(bersambung)

Jadual 4. (sambungan)

No. Item	Pengukuran DIF	DIF S.E.	Pengukuran DIF	DIF S.E.	DIF Contrast	Joint S.E.	t	d.f
	b. Inggeris		b. Melayu				Perbezaan	
13	-0.71	0.11	-0.53	0.11	-0.18	0.16	-1.15	995
14	-0.88	0.11	-1.1	0.12	0.22	0.17	1.3	995
15	0.14	0.1	0.35	0.1	-0.22	0.14	-1.54	995
16	0.77	0.1	0.91	0.1	-0.14	0.14	-0.95	995
17	-0.16	0.1	-0.16	0.1	0	0.15	0.03	995
18	1.04	0.1	1.09	0.1	-0.05	0.14	-0.34	995
19	1.24	0.1	1.43	0.1	-0.18	0.15	-1.26	994
20	-0.31	0.11	-0.54	0.11	0.23	0.15	1.5	995
21	0.57	0.1	0.51	0.1	0.06	0.14	0.41	995
22	0.32	0.1	0.51	0.1	-0.19	0.14	-1.37	995
23	0.01	0.1	-0.06	0.1	0.07	0.15	0.46	995
24	-0.65	0.11	-0.33	0.11	-0.32	0.15	-2.08*	995
25	-0.03	0.1	0.15	0.1	-0.18	0.14	-1.27	995
26	-0.27	0.1	-0.88	0.12	0.61	0.16	3.87*	995
27	0.61	0.1	0.68	0.1	-0.07	0.14	-0.49	995
28	-0.42	0.11	-0.21	0.1	-0.22	0.15	-1.45	995
29	-1.52	0.13	-1.47	0.14	-0.06	0.19	-0.3	995
30	1.31	0.1	1.19	0.1	0.12	0.15	0.8	995

DIF positif menunjukkan item bahasa Inggeris lebih sukar.

* menunjukkan DIF

Kebolehpercayaan ujian versi bahasa Melayu dapat ditingkatkan dengan mengeluarkan item 9, item 22 dan item 27. Untuk ujian versi bahasa Inggeris kebolehpercayaan meningkat apabila item 7, item 9 dan item 27 dikeluarkan. Item 9 dan item 27 merupakan item sepunya yang dapat meningkatkan kebolehpercayaan bagi kedua-dua versi ujian yang bahasa berlainan.

Jadual menunjukkan hanya dua item sahaja yang menunjukkan DIF berdasarkan kriteria $t \geq 1.96$ pada aras $p \leq 0.05$. Item-item yang ditandakan sebagai DIF ialah item 24 dan item 26 kerana nilai t masing-masing adalah -2.08 dan 3.87 .

Di samping itu, perbezaan di antara parameter kesukaran item yang ditentukan dengan WINSTEPS menunjukkan 14 item adalah lebih sukar dijawab dalam bahasa Melayu (perbezaan DIF adalah negatif), iaitu item 7, 9, 10, 13, 15, 16, 18,

19, 22, 24, 25, 27, 28 dan 29. Untuk baki 16 item yang mempunyai perbezaan DIF positif, item adalah lebih sukar dalam bahasa Inggeris. Berdasarkan jadual ini, kesimpulan yang dapat diperolehi ialah kedua-dua versi ujian ini adalah agak setara kerana hanya dua item yang menunjukkan DIF.

PERBANDINGAN PENCAPAIAN PELAJAR DALAM UJIAN VERSI BAHASA INGGERIS DAN BAHASA MELAYU

Perbandingan pencapaian pelajar dalam ujian yang berlainan bahasa iaitu bahasa Inggeris dan bahasa Melayu dapat ditunjukkan dengan menggunakan statistik deskriptif, graf bar dan taburan skor-z.

Jadual 5. Perbandingan pencapaian pelajar dengan statistik deskriptif

	Ujian versi bahasa Melayu	Ujian versi bahasa Inggeris
Skor minimum	4 (13.30%)	2 (6.70%)
Skor maksimum	30 (100.00%)	30 (100.00%)
Skor median	19 (63.30%)	18 (60.00%)
Skor min	18.746 (62.50%)	17.838 (59.50%)
Sisihan piawai	5.473	5.946
Varian	29.949	35.36
Min indeks diskriminasi	0.435	0.463
Min kesukaran item	0.625	0.595
<i>Skewness</i>	-0.292	-0.325
KR-20	0.820	0.845
Bilangan pelajar	500	500

Berdasarkan statistik deskriptif, didapati bahawa skor minimum untuk versi bahasa Inggeris dan versi bahasa Melayu masing-masing ialah 2 dan 4. Manakala skor maksimum untuk kedua-dua versi pula adalah sama, iaitu 30. Skor median dan skor min untuk versi bahasa Melayu adalah lebih tinggi berbanding dengan versi bahasa Inggeris. Skor median untuk versi bahasa Melayu ialah 19 dan versi bahasa Inggeris ialah 18. Sementara skor min versi bahasa Melayu ialah 18.746 adalah lebih tinggi sedikit berbanding dengan untuk versi bahasa Inggeris, iaitu 17.893. Secara keseluruhannya, pencapaian pelajar yang menjawab ujian versi bahasa Melayu adalah lebih baik daripada pelajar yang menjawab versi bahasa Inggeris. Nilai sisihan piawai dan varian versi bahasa Inggeris pula adalah lebih besar berbanding versi bahasa Melayu. Ini sepadan dengan nilai *skewness*, yang lebih besar untuk ujian versi bahasa Inggeris. Walau bagaimanapun, disebabkan nilai negatif yang wujud adalah agak dekat kepada nilai sifar, maka pencapaian

pelajar dalam kedua-dua versi ujian adalah cenderung kepada taburan normal. Namun, perbandingan antara min kesukaran item membuktikan bahawa ujian versi bahasa Melayu adalah lebih senang ($p = 0.625$) berbanding dengan versi bahasa Inggeris ($p = 0.595$). Untuk tujuan diskriminasi, ujian versi bahasa Inggeris merupakan ujian yang lebih baik.

Jadual 6 menunjukkan taburan skor pelajar yang telah ditukar kepada skor-z dan skor-z (ternormal). Secara umumnya bilangan pelajar yang mendapat skor-z nilai positif adalah lebih ramai untuk mereka yang mengambil ujian versi bahasa Melayu berbanding dengan yang mengambil ujian versi bahasa Inggeris.

Perbandingan dalam bentuk jadual tidak menunjukkan perbezaan yang ketara. Untuk memberi perbandingan yang lebih menyeluruh, graf garisan berdasarkan taburan skor-z dilukis seperti dalam Rajah 1.

Secara keseluruhan, graf garisan menunjukkan bahawa keputusan untuk ujian versi bahasa Melayu adalah lebih baik daripada ujian versi bahasa Inggeris. Antara skor-z bernilai -0.8 hingga 0.5 , lebih ramai pelajar yang menjawab ujian versi bahasa Melayu memperoleh skor dalam julat ini berbanding dengan pelajar yang menjawab ujian versi bahasa Inggeris.

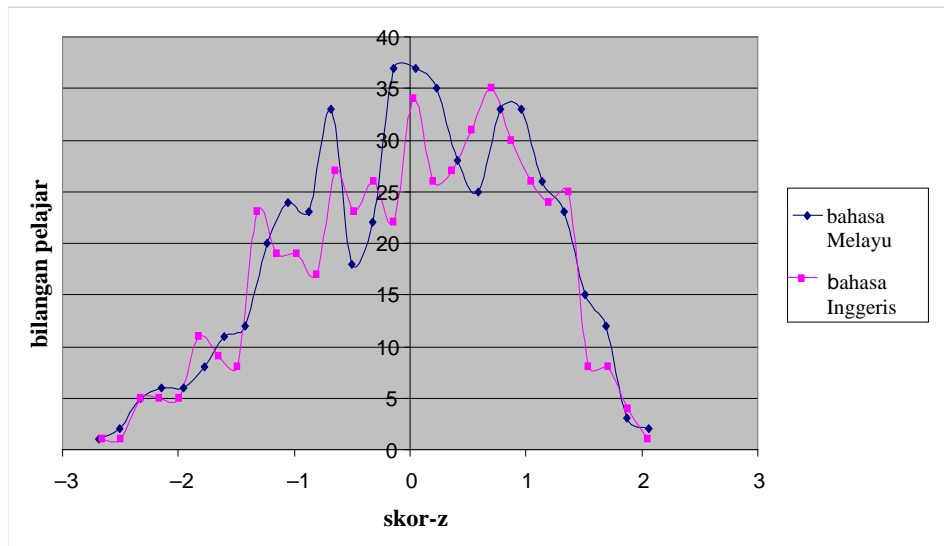
Jadual 6. Taburan markah pelajar untuk ujian versi bahasa Melayu dan ujian versi bahasa Inggeris

Skor	Kumpulan bahasa Melayu			Kumpulan bahasa Inggeris		
	Skor-z	Peratus longgokan	Skor-z ternormal	Skor-z	Peratus longgokan	Skor-z ternormal
2				-2.66	0.2	-3.09
3				-2.5	0.4	-2.74
4	-2.69	0.2	-3.09	-2.33	1.4	-2.36
5	-2.51	0.6	-2.65	-2.16	2.4	-2.07
6	-2.33	1.6	-2.29	-1.99	3.4	-1.89
7	-2.15	2.8	-2.01	-1.82	5.6	-1.69
8	-1.96	4	-1.82	-1.65	7.4	-1.51
9	-1.78	5.6	-1.66	-1.49	9	-1.39
10	-1.6	7.8	-1.49	-1.32	13.6	-1.21
11	-1.42	10.2	-1.34	-1.15	17.4	-1.01
12	-1.23	14.2	-1.16	-0.98	21.2	-0.86

(bersambung)

Jadual 6. (sambungan)

Skor	Kumpulan bahasa Melayu			Kumpulan bahasa Inggeris		
	Skor-z	Peratus longgokan	Skor-z ternormal	Skor-z	Peratus longgokan	Skor-z ternormal
13	-1.05	19	-0.97	-0.81	24.6	-0.74
14	-0.87	23.6	-0.79	-0.65	30	-0.6
15	-0.68	30.2	-0.61	-0.48	34.6	-0.45
16	-0.5	33.8	-0.46	-0.31	39.8	-0.32
17	-0.32	38.2	-0.35	-0.14	44.2	-0.2
18	-0.14	45.6	-0.2	0.03	51	-0.01
19	0.05	53	-0.01	0.2	56.2	0.09
20	0.23	60	0.16	0.36	61.6	0.22
21	0.41	65.6	0.32	0.53	67.8	0.37
22	0.59	70.6	0.47	0.7	74.8	0.56
23	0.78	77.2	0.64	0.87	80.8	0.76
24	0.96	83.8	0.85	1.04	86	0.97
25	1.14	89	1.09	1.2	90.8	1.19
26	1.33	93.6	1.35	1.37	95.8	1.49
27	1.51	96.6	1.65	1.54	97.4	1.82
28	1.69	99	2.01	1.71	99	2.09
29	1.87	99.6	2.45	1.88	99.8	2.51
30	2.06	100	2.87	2.05	100	3.09



Rajah 1. Graf garisan untuk pencapaian pelajar dalam ujian versi bahasa Inggeris dan bahasa Melayu dengan menggunakan z-skor.

PERBINCANGAN

Hasil dapatan kajian ini menunjukkan bahawa dari segi indeks diskriminasi, item versi bahasa Inggeris dan item versi bahasa Melayu tidak menunjukkan perbezaan yang ketara. Daripada analisis data, hanya tiga item menunjukkan perbezaan indeks diskriminasi yang lebih besar daripada 0.1, iaitu 0.18. Tambahan pula, hampir semua item mempunyai ciri diskriminasi yang baik, iaitu melebihi nilai 0.2 kecuali dua item untuk ujian versi bahasa Melayu dan satu item untuk versi bahasa Inggeris. Item yang mempunyai diskriminasi yang rendah adalah hasil daripada item yang terlalu mudah (item 6) atau item yang terlalu susah (item 9).

Bagi perbandingan aras kesukaran ujian pula, hampir semua item mempunyai perbezaan yang kurang daripada 0.1 kecuali item 26 yang mempunyai perbezaan yang paling besar dengan nilai 0.14. Item tersebut didapati lebih senang dalam versi bahasa Inggeris. Penentukur parameter menggunakan Model Rasch juga menyokong dapatan bahawa item 26 adalah lebih mudah dalam versi bahasa Inggeris.

Item yang paling susah iaitu item 9, dikenalpasti menggunakan statistik klasik dan juga Model Rasch. Hasil daripada itu, item 9 juga mempunyai indeks diskriminasi yang paling rendah, iaitu 0.18. Apabila pengkaji meneliti item

tersebut bersama-sama dua orang guru Sains, mereka mencadangkan sebab pencapaian yang rendah ialah pelajar tidak mengetahui punca yang menyebabkan hujan asid.

Dari segi kebolehpercayaan, item versi bahasa Inggeris adalah agak setara dengan item versi bahasa Melayu. Kebolehpercayaan untuk item versi bahasa Melayu ialah 0.82 manakala kebolehpercayaan untuk item versi bahasa Inggeris ialah 0.845. Nilai kebolehpercayaan adalah agak tinggi (melebihi 0.80) dan tidak berubah dengan ketara apabila item-item dikeluarkan daripada mana-mana versi ujian.

Keputusan analisis DIF hanya mengenalpasti dua item sahaja yang berfungsi berbeza apabila ditadbirkan dalam bahasa yang berlainan. Item DIF adalah item 24 dan item 26. Item 24 dan item 26 menunjukkan DIF kerana nilai t adalah masing-masing 2.08 dan 3.87 untuk $t \geq 1.96$ pada aras $p < 0.05$. Walau bagaimanapun, sekiranya kriteria yang lebih ketat dipilih dengan $t > 2.58$ pada aras $p < 0.01$, hanya satu item yang menunjukkan DIF iaitu item 26. Keputusan ini sejajar dengan analisis regresi logistik yang mengesan item 26 sahaja sebagai item DIF. Keputusan regresi logistik akan sama dengan keputusan Model Rasch jika kriteria yang digunakan adalah pada aras $p < 0.01$. Setelah meneliti item 26, didapati bahawa aras kesukaran item 26 dalam versi bahasa Melayu adalah jauh lebih tinggi daripada aras kesukaran versi bahasa Inggeris. DIF yang wujud mungkin disebabkan soalan item 26 untuk versi bahasa Inggeris adalah sukar difahami oleh pelajar berbanding dengan versi bahasa Melayu.

Dengan membandingkan pencapaian pelajar dalam kedua-dua versi ujian, didapati bahawa tidak terdapat perbezaan yang signifikan secara keseluruhannya. Walau bagaimanapun, dengan merujuk kepada kekerapan untuk setiap markah, pencapaian pelajar dalam ujian versi bahasa Melayu adalah lebih baik daripada ujian versi bahasa Inggeris. Daripada perbandingan pencapaian pelajar secara deskriptif, markah minimum untuk ujian versi bahasa Melayu ialah 4 berbanding dengan ujian versi bahasa Inggeris iaitu 2. Markah min untuk ujian versi bahasa Melayu pula adalah lebih tinggi, 18.746 berbanding dengan ujian versi bahasa Inggeris, iaitu 17.838.

KESIMPULAN

Dapatan kajian menunjukkan item untuk versi bahasa Inggeris adalah setara dengan item versi bahasa Inggeris berdasarkan analisis kebolehan diskriminasi, aras kesukaran, kebolehpercayaan, serta keberbezaan fungsi item (DIF). Selepas menentukan kesetaraan kedua-dua item dalam ujian, perbandingan pencapaian untuk ujian versi yang berlainan bahasa dapat ditentukan. Hasil perbandingan

menunjukkan bahawa secara keseluruhannya perbezaan markah pelajar untuk kedua-dua versi ujian adalah tidak ketara. Walau bagaimanapun, hasil kajian menunjukkan pencapaian pelajar dalam ujian versi bahasa Melayu adalah lebih baik berbanding dengan ujian versi bahasa Inggeris.

Dapatan ini memberi implikasi bahawa pentaksiran Sains perlu diteruskan dalam dwibahasa untuk kumpulan pelajar tertentu. Akomodasi demikian dapat membaiki kesahan ujian dan memastikan ujian membekalkan maklumat sebenar tentang kebolehan pelajar dalam sains, terutama sekali pelajar yang lemah dalam bahasa Inggeris. Namun, penggunaan dwibahasa mungkin lebih menguntungkan kumpulan yang tidak disasarkan, iaitu mereka yang cekap dalam kedua-dua bahasa Inggeris dan bahasa Melayu.

BIBLIOGRAFI

- Abedi, J. (2006). Psychometric issues in the ELL assessment and special education Eligibility. *Teachers College Record*, 108(11), 2282–2303.
- Allalouf, A., Hambleton, R. K. dan Sienchi, S. G. (1999). Identifying the causes of differential item functioning in translated verbal items. *Journal of Educational Measurement*, 36, 185–198.
- Angoff, W. H. dan Cook, L. L. (1988). *Equating the scores of the Prueba de Aptitud Academica and the Scholastic Aptitude Test* (Report No. 88–2). New York: College Entrance Examination Board.
- Bond, T. G. dan Fox, C. M. (2001). *Applying the Rasch model: Fundamental measurement in the human sciences*. Mahwah, NJ: Erlbaum.
- Budgell, G. R., Raju, N. S. dan Quartetti, D. A. (1995). Analysis of differential item functioning in translated assessments. *Applied Psychological Measurement*, 19, 309–321.
- Crocker, L. dan Algina, J. (1986). *Introduction to classical and modern test theory*. New York: Holt, Rinehart and Winston.
- Ebel, R. L. dan Frisbie, D. A. (1986). *Essentials of educational measurement*. Englewood Cliffs, NJ: Prentice-Hall.
- Geisinger, K. F. (1994). Cross-cultural normative assessment: Translation and adaptation issues influencing the normative interpretation of assessment instruments. *Psychological Assessment*, 6, 304–312.

- Hambleton, R. K. (1994). Guidelines for adapting educational and psychological tests: A progress report. *European Journal of Psychological Assessment*, 10, 229–244.
- Kolen, M. J. dan Brennan, R. L. (1995). *Test equating: Methods and practices*. New York: Springer.
- Linacre, J. M. (2005). *A user's guide to WINSTEPS*. [Computer Program]. Chicago: Winsteps.
- Lord, F. M. (1980). *Application of item response theory to practical testing problems*. Hillsdale, NJ: Erlbaum.
- Olmedo, E. L. (1981). Testing linguistic minorities. *American Psychologist*, 36, 1078–1085.
- Prieto, A. J. (1992). A method for translation of instruments to other languages. *Adult Education Quarterly*, 43, 1–14.
- Stansfield, C. W. (1996). Content assessment in native language: *Eric Digest*. <http://www.ericdigest.org/1998-1/content.htm> (dicapai pada 24 Mei 2007).
- The Star. (2005). Students Still Prefer to Use Bahasa. <http://www.apps2.emoe.gov.my/tayang.php?laman=bacaakhbar&bhs=en&id=1065> (dicapai pada 23 Desember 2005).
- Van de Vijver, F. J. R. dan Leung, K. (1997). *Methods and data analysis for cross cultural research*. Thousand Oaks, CA: SAGE Publication.
- Woldbeck, T. (April, 1998). *Basic concepts in modern methods of test equating*. Kertas kerja dibentangkan di Annual Meeting of the Southwest Psychological Association, New Orleans, LA.
- Wood, D. A. (1960). *Test construction: Development and interpretation of achievement tests*. Columbus, OH: Charles E. Merrill Books, Inc.
- Zhu, W–M. (1998). Test equating: What, why, how? *Research Quarterly for Exercise and Sport*, 1, 11–23.