Jurnal Pendidikan Malaysia 35(1)(2010): 1-10

Differential Item Functioning in Malaysian Generic Skills Instrument (MyGSI)

Siti Rahayah Ariffin, Rodiah Idris & Noriah Mohd Ishak

ABSTRACT

Apart from good academic performance, generic skills are necessary for students in order to be more successful and to excel as practitioners in academic fields, work and careers. The development of the Malaysian Generic Skills Instrument (MyGSI) is based on the Malaysian Qualification Framework which applies cognitive, behaviorism and social theories. The purpose of this study is to detect Differential Item Functioning (DIF) in MyGSI based on gender, types of program and race. Data was obtained from 1,262 undergraduate students randomly chosen from 12 faculties at Universiti Kebangsaan Malaysia (UKM). The MyGSI used consists of 13 constructs and 102 items on a 4-point likert scale. In addition, Winsteps software version 3.64.2 has been used for data analysis. The findings had detected five misfitting items; one item was distinguished as DIF based on types of programme and 12 items were detected as DIF based on race. The final analysis using the Rasch's model has dropped 15 items and maintained 87 items that were legitimate and reliable to gauge 13 constructs in MyGSI. This MyGSI, free from DIF, could be used to obtain Higher Learning Institutions (HLI) students' profile in a justly manner. In short, it could be used as an indicator to increase students' generic skills during their study in the university.

Keywords: Generic skills, differential item functioning, Rasch Measurement Model, item polarity, item misfit

ABSTRAK

Kemahiran generik adalah kemahiran yang diperlukan oleh pelajar selain akademik untuk menjadi lebih berjaya dan cemerlang sebagai pengamal di dalam bidang akademik, pekerjaan dan kehidupan. Pembinaan Instrumen Kemahiran Generik Malaysia (MyGSI) adalah berdasarkan Kerangka Kelayakan Malaysia yang mengaplikasikan teori kognitif, behaviorisme dan sosial. Tujuan kajian ini adalah untuk mengesan Kebezaan Kefungsian Item (DIF) berdasarkan faktor gender, aliran pengajian dan bangsa. Kajian ini melibatkan 1,262 pelajar prasiswazah yang dipilih secara rawak daripada 12 fakulti di Universiti Kebangsaan Malaysia (UKM). Instrumen MyGSI mengandungi 13 konstruk dan 102 item skala likert 4-mata. Perisian Winsteps versi 3.64.2 digunakan untuk analisis data. Hasil kajian menunjukkan sebanyak lima item dikesan misfit, satu item dikesan DIF berdasarkan aliran dan 12 item dikesan DIF berdasarkan bangsa. Analisis akhir menggunakan model Rasch telah mengugurkan 15 item dan mengekalkan 87 item yang sah dan boleh dipercayai untuk mengukur 13 konstruk dalam MyGSI. Instrumen MyGSI yang bebas DIF ini boleh digunakan untuk mendapatkan profil pelajar-pelajar Institusi Pengajian Tinggi (IPT) secara lebih adil dan saksama sebagai indikator bagi meningkatkan kemahiran generik pelajar sepanjang pengajian di universiti.

Kata kunci: Kemahiran generik, kebezaan kefungsian item, Model Pengukuran Rasch, polariti item, misfit item

Introduction

The evaluative process and quality assurance of Higher Learning Institution (HLI) graduates are beneficial. Even though it is difficult to develop, generic skills achievement of students in HLI could be determine (Hambur et al. 2002). Evaluating students' competency is very challenging. Several researchers who proposed logical procedures for the study of DIF could be mentioned along this line of research (Berk 1982; Cole 1981; Hambleton & Jones 1994; Scheuneman 1987; Shepard 1982; Title 1982). They recommend that aspects related to the composition and format of the items should be taken into consideration to avoid bias.

Generic skills are also essential for students to be successful in academic fields (Falk & Millar 2002; Hambur et al. 2002; Lublin 2003; Siti Rahayah et al. 2008a). They are integrated within the educational context of teaching and learning (Kearns 2001). The development of the Malaysian Generic Skills Instrument (MyGSI) is based on the Malaysian Qualification Framework (MQF) 2006 including cognitive, behaviorism and social theories. MQF stresses 8 domains of learning outcomes which consists of; (1) disciplinary knowledge (2) practical skills (3) social skills and responsibilities (4) values, attitudes and professionalism (5) communication, leadership and teamwork skills (6) critical thinking, problem-solving and scientific skills (7) information management and lifelong

learning skills (8) managerial and entrepreneurial skills (Sharifah Hapsah 2006).

The development of MyGSI is based on cognitive theory (Ausubel 1978). Learning happens when past learning experience affects subsequent learning performance. Hence, it is known as transferable skills (Kearns 2001). Meanwhile, behaviourism theory leads to behavioural changes. It covers personal needs and intentional behaviours. Skinner (1971) suggests that human being is moulded based on one's learning process and interaction with his environment. Therefore, cognitive theory and behaviourism are approaches in constructing MyGSI (Rodiah 2008b).

HLI is the most suitable place to develop students' generic skills (Allan & Clarke 2007; Ballantine & Larres 2007; Bennett et al. 2000; Biggs 2003; Jager & Nassimbeni 2005; Lizzio & Wilson 2004; Lublin 2003). HLI students must have the desire to form human capital with 'first-class mentality'. Hence, overall generic skills acquisition must be carried out by the university.

The objective of the study is to examine the psychometric characteristics of the MyGSI from various aspects, namely (1) reliability and validity; (2) difficulty of MyGSI items, and (3) the existence of DIF based on demographic factor measure, such as gender, types of program and race.

The basic unit of instrument measurement is an item. Item creation must be firm and equal to all participants. DIF refers to item with different functions to measure a construct. It is being administered to a group of diverse demographic of background but similar capability respondents. Hambleton et al. (1991) suggest an item detected by DIF is dissimilar in terms of functions in diverse subgroups. Therefore, DIF analysis procedure is designed to recognize items that do not mirror similar functions when administered to groups with parallel capability. Item functioning traits that are being compared in this research is item difficulty index.

Each item in the instrument must be tested on its suitability before it is administered to respondents. Osterlind (1989) states that item analysis involve studying items by critical observation in order to reduce measurement error. Thus, DIF analysis is used to determine items validity (Ackerman 1992). According to Camilli (1993), DIF and validity are interrelated. Berk (1982) explains DIF study is essential due to its most basic level in content analysis and inference. DIF is unwarranted since the test does not measure the same ability in distinct groups (Maller 2001).

DIF endorsement in instrument construction is an indicator of high reliability instrument. Angoff (1993) and Siti Rahayah et al. (2008) believe DIF affects instrument reliability. Three DIF endorsement methods are Mantel-Haenszel (Dodeen & Johanson 2001; Stoneberg 2004), Item Response Theory (Maller 2001; Lamprianou et al. 2002) and Rasch Models (Cauffman & MacIntosh 2006). This study uses Rasch Models to identify gender, types of programmes and race. Bond and Fox (2001, 2007) suggest three DIF indicators based on the studied groups which are (1) t value ± 2.0 ($t \geq +2.0 \leq -2.0$), (2) DIF Contrast ± 0.5 (DIF Contrast $\geq +0.5 \leq -0.5$), and (3) $p < 0.05$.

DIF analysis is directly done in academic and non-academic researches. Academically, DIF analysis is employed in Science, Mathematics, English, History and Economy subjects (Stoneberg 2004; Zwick & Ercikan 1989). DIF is used in instrument research, qualifying test, promotion, license and products granting (Dodeen 2004; Zieky 2002).

Gender, types of programmes and race are factors that need to be focused in this research because they are important features in Malaysian educational system (Siti Rahayah et al. 2008a). Zieky (1993) chooses popular subgroups in the community. Gibson and Harvey (2003); O'Neill & McPeek (1993); Siti Rahayah et al. (2008e) compare students' achievement based on gender. Their research has become references in different fields. DIF analysis on the item needs to be done to ensure its reliability. Malaysian educational system comprises of science and non-science streams. Item construction in measuring generic skills needs to consider both streams. This would measure students' capability and avoid bias item or DIF. Race DIF item verification is compulsory because the study involves three races: Malay, Chinese and Indian. Hence, the items are free from race DIF to ensure its justification.

Related DIF studies based on types of programmes, race, year and gender have been conducted by Siti Rahayah et al. (2008a, 2008b, 2008c). They are related to generic skills instrument construction but emphasize on three skills which are teamwork, communication and leadership. Sheppard et al. (2006) studied on Hogan Personality Inventory and have discovered 38.4% (53 out of 138 items) gender based DIF and 37.7% (52 from 138 items) racial based DIF (potentially biased more for Caucasians than Blacks). Cauffman and Maclntosh (2006) studies on Massachusetts Youth Screening Instrument detect some items contain gender and racial DIF.

Rasch Models

Rasch (1960) created distinguishable but inactive "Rasch Models". They reflect the probability between the level of inactive trait (or person ability or measure) and the measurement items (item location or difficulty). Rasch analysis meets adequate standards in statistics. It is designed to aid deciding on intertwined data matters (Wright & Masters 1982).

According to Linacre (2002), Rasch models are the transformation of ordered qualitative observations into linear measures. The models function based on Winsteps which are: (1) The dichotomous model: $\log(Pni1 / Pni0) = Bn - Di$. (2) The polytomous "Rating Scale" model: $\log(Pnij / Pni(j-1)) = Bn - Di - Fj$. (3) The polytomous "Partial Credit"

model: log(Pnij/ Pni(j-1)) = Bn - Di - Fij = Bn - Dij. (4) The polytomous "Grouped response-structure" model: log(Pnij/ Pni(j-1)) = Bn - Dig - Fgj. Where: Pnij is the probability person n encountering item i is observed in category j, Bn is the "ability" measure of person n, Di is the "difficulty" measure of item i, the highest and lowest categories point of the item are probable. Fj is the "calibration" measure of category j relative to category j-1, the point categories j-1 and j are probable relative to the measure of the item. A higher ability respondent would score high in probability of endorsement of an item than a lower ability respondent (Bond & Fox 2001). Difficult items are tough to score due to lower probability of endorsement (Smith 2000).

Rasch analysis converts raw data from scores to logits. The logits are compared to a linear model to find its odds of success. It is subjected to floor and ceiling effects; ranging from 0 to 1. The logs denote the natural log of an odds ratio (Andrich 1978; Smith 2000). Bond and Fox (2001) state reliability is measured by the ability of the scale to locate the level of the attribute. When diverse populations are employed to assess the same construct in changed environments, the same ability should be produced (Nunnally & Bernstein 1994). American Psychological Association (1985) defines validity as the extent to which meaningful inferences can be made from the measurement. Two elements of validity are criterion and construct. Criterion-related validity scrutinizes the measure ability to predict an outcome. Construct validity observes whether the items used in the measure reflect the concept, construct, or dimension being measured.

Classical test theory (CTT) suggests point-biserial correlations should be 0.3, 0.4 or better to determine the construct validity. According to Linacre (2007), point-biserial (or point-measure) correlations should be positive. Every item should add a significant approach to the construct/concept (Bold & Fox 2007). The suitable item is calculated by the mean-square residual fit statistics (Bond & Fox 2001). Fit statistics expected value is 1.0, and ranges from 0 to infinity. Deviations denote lack of fit between the items and the model. The lower values than expected could be interpreted as item redundancy or overlap.

McNamara (1996) defines fit as the observed responses for individual items which strengthen the general pattern displayed in the matrix. Infit statistics-Information-weighted fit statistics are perceptive to unanticipated behaviors affecting responses to items (Linacre 2002). Outfit statistics-outlier-sensitive fit statistics are sensitive to unexpected behavior of the person on items (Linacre 2002). Item fit statistics-MNSQ (mean squares) expected value is 1.0. MNSQ > 2.0 degrades the measurement system. Value $1.5 <$ MNSQ ≤ 2.0 is unproductive for measurement construction. Value 0.5 MNSQ ≤ 1.5 is productive for measurement. MNSQ < 0.5 is less productive for measurement (Linacre 2002; Smith 2000). Bond and Fox (2007) suggest item mean square for Infit and Outfit scale (likert/survey) ranges from 0.6 to 1.4.

DIF is functioning when one group of respondents managed to score higher than another group on the same item. This could represent: (1) One group achieves its natural "attitude/ability" level, the other performs better than usual; (2) One group performs its usual "attitude/ability" level, the other performs shoddier than usual; (3) The item has its usual difficulty for one group, but is more difficult than usual for the other; (4) The item has its usual difficulty for one group, but is simple than usual for the other. Smith (2000) believes item parameters should be similar across populations.

Methodology

This study is conducted by using quantitative approach. A total of 16,189 UKM undergraduate students have been the target population of the study. A clustered random sampling based on 12 learning faculties was used in this study. A total of 1,262 students comprising of 377 males, 885 females, out of which 665 are science students, 597 from non-science backgrounds were selected as a sample for the study. Students were also representative of the various races in Malaysia, namely the Malay, Chinese, and Indian.

MyGSI is used to measure students' generic skills. Students circled their agreement to the items using the 4-point Likert-type response categories (1 = disagree, 2 = less agree, 3 = agree, and 4 = strongly agree). MyGSI is used to gauge 13 generic skills with 102 items which comprises of (1) Social Responsibility - SocialR (7 items), (2) Environment Awareness - EnvironmentA (5 items), (3) Ethics, Morals and Professionalism - EthicMP (5 items), (4) Spiritual - Spirit (6 items), (5) Communication - Com (13 items), (6) Leadership - Leader (10 items), (7) Teamwork - Team (9 items), (8) Critical Thinking and Problem Solving - CriticalTPS (10 items), (9) Information Technology and Communication - ICT (7 items), (10) Lifelong Learning - LifelongL (12 items), (11) Globalization - Global (7 items), (12) Entrepreneurship - Entrepreneur (6 items) and (13) Managerial - Manager skills (5 items). Data is analyzed using Winstep's software 3.62.4 (Linacre 2007) based on Rasch's model (Rasch 1960) to check on item fit and DIF.

Results and Discussion

The total number of respondents is 1,262. There were 885 respondents who were female (70.1%) and 377 of them were male (29.9%). The total number of Malay students was 927 (73.5%), Chinese 249 (19.7%), Indian 52 (4.1%) and others 34 (2.7%). There are 665 science stream students (52.7%) and 597 non-science students (47.3%).

Respondent's reliability reflects an equivalent to Cronbach Alpha or KR20 measurement (Master 1982; Wright & Master 1982). Table 1 shows respondents' and item reliability index of 13 MyGSI constructs.

Acceptable respondents' reliability index is from 0.80 to 1.00 which indicates positive feedback. In addition, it stands at the same par with the consistency level of arrangement. In other words, it is above the logit scale for different sets of item agreement which measures the same construct. The respondents' separation index that could be detected in this study is from 2 to 3. Therefore, it denotes that respondents' reliability is beyond MyGSI items which is within 2 to 3 endorsement level. Acceptable value of respondents' reliability index is ≥ 0.8 and of separation index is ≥ 2.0 (Bond & Fox 2001, 2007; Linacre 2002).

Table 1 illustrates item reliability index is from 0.85 to 1.0. Wright & Masters (1982) claim the value is positive because it is near to 1.0. Hence, MyGSI item repetition prediction is also high if it is being administered to other groups of respondents with similar capability (Wright & Masters 1982). Item separation index is from 2 to 17. Statistically, it reflects MyGSI item could be divided from 2 until 17 strata or endorsement level. Moreover, this provision connotes that the items are 2 to 17 times more distributed from mean square error. The respondents' reliability index and item ≥ 0.8 is acceptable (Fox & Jones 1998; Bond & Fox 2001). In addition, the separation index of ≥ 2.0 is also acceptable (Fox & Jones 1998).

TABLE 1. Reliability Value and Respondent Separation Index and Item

No	Construct	Item (N)	Respondent		Item	
			Reliability	Separation	Reliability	Separation
1.	SocialR (s1)	7	0.84	2.28	0.96	4.75
2.	EnvironmentA (a2)	5	0.78	1.87	1.00	15.58
3.	EthicMP (et3)	5	0.81	2.05	0.85	2.40
4.	Spiritual (r4)	6	0.81	2.06	0.99	9.85
5.	Communication (c5)	13	0.87	2.62	0.99	8.68
6.	Leadership (L6)	10	0.86	2.47	0.97	5.27
7.	Teamwork (t7)	9	0.84	2.28	0.99	11.33
8.	CriticalTPS (k8)	10	0.89	2.86	0.97	6.20
9.	ICT (it9)	7	0.79	1.94	1.00	17.35
10.	LifelongL (LLL10)	12	0.89	2.78	0.95	4.30
11.	Globalization (g11)	7	0.85	2.34	0.99	13.08
12.	Entrepreneur (u12)	6	0.87	2.53	0.97	5.88
13.	Managerial (p13)	5	0.90	2.93	0.99	8.27

Table 2 shows a summary of point measure correlation (PTMEA CORR) for 102 items in MyGSI. All items show positive value with index > 0.30. Minimum PTMEA CORR index is 0.36 of item g87 (globalization) and maximum index is 0.74 of item p102 (managerial). According to Bond & Fox (2001) the positive value of PTMEA CORR proves measuring items that are to be measured need to be carefully constructed. Therefore, all items in MyGSI are measuring 13 generic skill constructs. This analysis is the basic step to gauge the validity of construct used to build and validate MyGSI instrument. PTMEA CORR index will increase if misfitting items are dropped from cluster item measurement.

TABLE 2. Point Measure Correlation (PTMEA CORR) of MyGSI Construct

ENTRY NUMBER	RAW SCORE	COUNT	MEASURE	MODEL S.E.	INFIT		OUTFIT		PTMEA CORR.	ITEM
					MNSQ	ZSTD	MNSQ	ZSTD		
87	3925	1262	1.58	0.04	1.96	9.9	2.54	9.9	0.36	g87
102	4416	1262	0.78	0.04	0.85	-4.2	0.87	-3.4	0.74	p102

Figure 1 shows numbers of respondents and the difficulty of items capability hierarchy above a logit scale. The results confirm all items are scattered and pointing towards the capability level of respondents' diversity. The ranking of respondents with high capability (easily agree) is above the scale, whilst the ranking of lower respondents (difficult to agree) is below the scale. The item which is difficult to be agreed upon is g87 with difficulty to be measured is 1.58 logit on the top scale, whilst the simplest item to be agreed upon is item a12 with measurement of -1.18 logit on the lower scale. The item which is difficult could be answered by respondents with high capability, whilst easy items could be answered by respondents with high and low ability (Linarce 2007). Overlapping items measure different elements with different levels of difficulty (Bond & Fox 2001, 2007).

INPUT:	1262	person	102	items	MEASURED:	1262	person	102	items	4	CATS
		individus - MAP - items									
		<more>|<rare>									
4		.	+								
		.	|								
		.	|								
		.	|								
		.#	|								
3		.#	T +								
		.#	|								
		.#	|								
		##	|								
		.###	|								
2		.###	+								
		.###	S |								
		.#####	|	g87							
		.######	|	it70							
		.######	|	g85							
1		.##########	+ T	c25							
		.##########	|	c28	p102						
		.###########	M | S	c24	g89	g90	p100	p101	r18	r20	t48
				u92								
		.##########	|	a11	c29	g86	g88	k56	k64	k65	p99
				r21								
		.###########	|	c26	c27	c30	c32	c33	c34	c35	et13
				et15	g91	it66	it68	it72	k57	k59	k62	
				k63	p98	u94	u96	u97				
0		.########	+ M	c31	et14	et17	it69	k58	k61	l39	l42	
				lll74	lll76	lll78	r19	r23	u93			
		.########	|	a10	c36	et16	k60	l43	l44	l46	lll73	
				lll75	lll77	lll79	lll81	lll82	lll83	t47	t55	
				u95								
		.########	|	l40	l41	l45	lll84	s2	s7	t49	
		.######	S | S	a9	it67	l37	l38	lll80	r22	s1	s3
				t50	t51	t53	t54					
		.####	|	s4	s5	t52					
-1		.###	+ T	a8	it71	s6					
		.##	|	a12							
		.#	|								
		.	|								
		.	T |								
-2		.	+								
		.	|								
			|								
EACH '#' IS 9.											

FIGURE 1. Map of MyGSI Respondent-Item

Table 3 shows the item fit index (infit/outfit MNSQ) of 102 items in MyGSI. The examination result of infit/outfit MNSQ shows 5 items demonstrating values infit/outfit MNSQ that are well above 1.40 logit, namely item a12 = 1.48/1.45 logit, a8 = 1.54/1.54 logit, g87 = 1.96/2.54 logit, r22 = 2.21/2.12 logit and s1 = 1.53/1.54 logit. Bond & Fox (2001) explain that the acceptable range is between 0.6 to 1.4 logit. Higher value of 1.4 logit shows items that are not homogeneous with other items within one measurement scale. Items which are less than 0.6 logit show overlapping items with the others. Items which need further attention or those items that have been dropped are items a12, a8, g87, r22 and s1.

Analysis has been carried out to study the existence of Gender Differential Item Functioning (GDIF) in the instrument used. To analyse GDIF, Winstep performs two-tailed t-test to investigate the significant difference between two index difficulties. The confidence level is 95% and the level of t critical value rests with value 2.0 for all DIF analysis. In addition, GDIF Contrast index is used to show the difference of gap confirmation level for each item when males and females are being compared.

TABLE 3. Item Statistics: Misfit Order

ENTRY NUMBER	RAW SCORE	COUNT	MEASURE	INFIT		OUTFIT		PTMEA CORR.	ITEM
				MNSQ	ZSTD	MNSQ	ZSTD		
12	5400	1262	-1.18	1.48	9.9	1.45	8.5	0.4	a12
8	5325	1262	-1	1.54	9.9	1.54	9.9	0.38	a8
87	3925	1262	1.58	1.96	9.9	2.54	9.9	0.36	g87
22	5146	1262	-0.58	2.21	9.9	2.12	9.9	0.49	r22
1	5122	1262	-0.53	1.53	9.9	1.54	9.9	0.37	s1

According to Lai Eton (2002), value of 0.5 logits DIF contrast would be vital for likert scale. Meanwhile, Wright and Panchalakesan in Pallant and Tennant (2007) argue that the size of GDIF which is less than 0.5 logits is considered unimportant (DIF negligible). A negative index of GDIF Contrast means that the item is easier to be confirmed by males while a positive index item is easier to be confirmed by female respondents. DIF Measurement is the difficulty index of this item for this group, with other elements held constant.

The analysis demonstrates that 28 items (27%) from 102 items in MyGSI show the significance of GDIF in value t ≥ 2.0 logit. The GDIF Contrast (≥ 0.5 logit) shows that 28 items do not show serious GDIF because the GDIF index shows less than 0.5 logit. As such, it is proposed the 28 items could be maintained.

The analysis of DIF based on types of program denotes 32 items (31%) out of 102 items in MyGSI show DIF significant (t ≥ 2.0 logit). The DIF Contrast (≥ 0.5 logit) only demonstrates at item s1. As such it is proposed that item s1 ("I am responsible to myself and others") be dropped because t-value is 6.6 logit and DIF contrast is - 0.61 (Table 4). Based on DIF measure (Difficulty measure), item s1 is easier to be agreed by males but is difficult to be agreed by females.

TABLE 4. Differential Item Functioning Analysis Based on Stream – MyGSI

GROUP (SCIENCE)	DIF MEASURE (DIFFICULTY MEASURE)	GROUP (NON SCIENCE)	DIF MEASURE (DIFFICULTY MEASURE)	DIF CONTRAST (DIF SIZE)	t	ITEM
1	-0.83	2	-0.22	-0.61	-6.6	s1

Table 5 shows 12 items of 102 items in MyGSI related to race and demonstrate the significance of DIF. All of the 12 items (12%) with DIF significance are based on t value; more than 2.0 logit (t ≥ 2.0 logit) and DIF Contrast ≥ 0.5 logit. Consequently, examination results show 12 items (s1, a8, et13, c24, c27, k58, it68, lll76, lll77, g85, g86 and u97) in MyGSI are found compatible with DIF significant serious race. It is proposed that they can be dropped.

Item s1 ("I am responsible to myself and others") is easier to be agreed by Malay (DIF measure = -0.53 logit) and Chinese (DIF measure = -0.62 logit) but difficult to be agreed by Others (DIF measure = 0.19 logit). Item a8 ("I support any activities which are related to environment appreciation") is easier to be agreed by Others (DIF measure = -1.04 logit), Chinese (DIF measure = -0.82 logit) and Malay (DIF measure = -1.1) but difficult to Indian (DIF measure = 0.04 logit). Item et13 ("I act in a profesional manner during my working") is easier to be agreed by Malay (DIF measure = -0.03 logit) but difficult to Chinese (DIF measure = 0.47 logit), Indian (DIF measure = 0.53 logit) and Others (DIF measure = 0.83 logit). Item c24 ("I give feedback on issues presented at every level through the available channels of the university") is easier to be agreed by Malay (DIF measure = 0.63 logit) but difficult to Others (DIF measure = 1.14 logit).

Item c27 ("I can express my ideas clearly") is easier to be agreed by Indian (DIF measure = -0.35 logit) but difficult to Malay (DIF measure = 0.28 logit) and Chinese (DIF measure = 0.2 logit). Item k58 ("I can distinguish between main problem and the hidden objectives easily') is easier to be agreed by Indian (DIF measure = -0.53 logit) but difficult to Malay (DIF measure = 0.09). Item it68 ("I can use effective techniques during electronic research") is easier to be agreed by Chinese (DIF measure = 0.05 logit) but difficult to Indian (DIF measure = 0.71 logit). Item lll76 ("I can choose self-quality source efficiently") is esier to be agreed by Malay (DIF measure = 0.03 logit) and Chinese (DIF measure = -0.26 logit) but difficult to Indian (DIF measure = 0.53 logit). Item lll77 ("It is usually easy for me to receive new ideas") is easier to be agreed by Chinese (DIF measure = -0.55 logit) but difficult to Indian (DIF measure = -0.01 logit) and Others (0.06 logit).

Item g85 ("I can speak proficiently in English and I am comfortable when facing the masses") is easier to be agreed by Indian (DIF measure = 0.1 logit) but difficult to Malay (DIF measure = 1.27 logit); easier to be agreed by Indian (DIF measure = 0.1 logit) but difficult to Chinese (DIF measure = 0.74 logit); easier to be agreed by Chinese (DIF measure = 0.74 logit) but difficult to Malay (DIF measure = 1.27 logit); easier to be agreed by Indian (DIF measure = 0.1 logit) but difficult to Others (DIF measure = 0.78 logit). Item

TABLE 5. Differential Item Functioning Analysis Based on Race - MyGSI

GROUP 1-MALAY 2-CHINESE 3-INDIAN 4-OTHERS	DIF MEASURE (DIFFICULTY MEASURE)	GROUP 1-MALAY 2-CHINESE 3-INDIAN 4-OTHERS	DIF MEASURE (DIFFICULTY MEASURE)	DIF CONTRAST (DIF SIZE)	t	ITEM
1	-0.53	4	0.19	-0.72	-2.73	s1
2	-0.62	4	0.19	-0.81	-2.91	s1
3	0.04	4	-1.04	1.07	2.85	a8
2	-0.82	3	0.04	-0.86	-3.39	a8
1	-1.1	3	0.04	-1.14	-4.8	a8
1	-0.03	2	0.47	-0.5	-4.71	et13
1	-0.03	3	0.53	-0.56	-2.55	et13
1	-0.03	4	0.83	-0.86	-3.41	et13
1	0.63	4	1.14	-0.5	-2.01	c24
1	0.28	3	-0.35	0.63	2.53	c27
2	0.2	3	-0.35	0.54	2.09	c27
1	0.09	3	-0.53	0.62	2.44	k58
2	0.05	3	0.71	-0.66	-2.87	it68
1	0.03	3	0.53	-0.51	-2.28	lll76
2	-0.26	3	0.53	-0.79	-3.33	lll76
2	-0.55	3	-0.01	-0.53	-2.1	lll77
2	-0.55	4	0.06	-0.6	-2.13	lll77
1	1.27	3	0.1	1.17	5.03	g85
2	0.74	3	0.1	0.64	2.6	g85
1	1.27	2	0.74	0.53	5.17	g85
3	0.1	4	0.78	-0.68	-2.01	g85
1	0.56	3	-0.35	0.91	3.67	g86
2	0.28	3	-0.35	0.62	2.4	u97
1	0.25	3	-0.35	0.59	2.4	u97
3	-0.35	4	0.52	-0.87	-2.48	u97

g86 ("I can build meaningful sentence structures in writing") is easier to be agreed by Indian (DIF measure = -0.35 logit) but difficult to Malay (DIF measure = 0.56 logit). Item u97 ("I can assess the results of a proposal and the planning made") is easier to be agreed by Indian (DIF measure = -0.35 logit) but difficult to Chinese (DIF measure = 0.28 logit); easier to be agreed by Indian (DIF measure = -0.35 logit) but difficult to Malay (DIF measure = 0.25 logit); easier to be agreed by Indian (DIF measure = -0.35) but difficult to Others (DIF measure = 0.52 logit).

Research findings on MyGSI inspection have resulted in 102 positive items. This proves all items are measuring generic skills. MNSQ outfit/Infit analysis produces five misfit item (social responsibility, environmental care, spiritual and globalization items) based on Rasch models. The five misfit items decrease the overall item reliability. The exclusion of the items would increase the MyGSI reliability index. The misfit items are unnecessary because they are irrelevant in measuring generic skills.

Person-map item has clearly shows item g87 ("I have published my research article in English") is the hardest to be endorsed. This reflects UKM undergraduates students have never or lack in article publication. Only 17% of the students has endorsed g87. Meanwhile, item a12 ("The university must hold campaigns to cultivate appreciation of the environment as part of its annual programme") is the easiest to be endorsed. Such programmes and activities have been regularly carried out by the university. Almost 96% of the students endorsed the item. Both the hardest and easiest items to be endorsed are misfit. This is because the five identified misfit items are also difficult or easy items to be endorsed.

All 102 items are free from gender based DIF. The study contradicts Sheppard, Han, Colarelli and Dai (2006) studies on Hogan Personality Inventory which has discovered that 38.4% (53 out of 138 items) gender based DIF and 37.7% (52 from 138 items) race based DIF (potentially biased more for Caucasians than Blacks). Cauffman and Maclntosh (2006) studies on Massachusetts Youth Screening Instrument detect some items containing gender DIF.

Meanwhile, one item (social responsibility skills) contains types of program based DIF. There are 12 items detected with race DIF. Item s1 (social responsibility: "I am responsible to myself and others") was detected with types of program and race DIF. The item could be dropped because it is misfit. To be more exact, there are four items that need to be dropped which are a8, et13, g85 and u97. Item a8 (environmental awareness: "I support any activities which are related to environmental appreciation") is difficult to be endorsed by Indian but

easy by Malay and Chinese. Therefore, item a8 is a misfit. Item et13 (ethic, moral and profesionalisme: "I act in a profesional manner during my working") is difficult to be endorsed by the Chinese, Indian and others but not to Malay. Item g85 (globalization: "I can speak proficiently in English") is the hardest item to be endorsed by all races as detected by DIF. Hence, item g85 needs to be dropped. Meanwhile, item u97 (entrepreneurship: "I can assess the results of a proposal and the planning made") is easily endorsed by Indian but not by the Chinese and Malay.

Although the research has reflected five misfit items, it has no DIF significant. The detected DIF items do not influenced 13 generic skills constructs. All of the 13 constructs are free from gender DIF. Only social responsibilities construct is identified as types of program DIF. Spiritual, leadership, teamwork, and managerial skills constructs are freed from race DIF.

To summarize, there are 15 items or 15% (5 misfit items and 12 DIF items) that need to be dropped in the research. Such action would enhance the reliability and validity of MyGSI and increases item quality in measuring generic skills. The instrument validity would also be affected by unjust instrument matters. Angoff (1993) states DIF items influence its validity score and have serious effects on the respondents.

The study finding is parallel to studies by Covic et al. (2007); Hambleton & Rogers (2000); Higgins (2007); Lamprianao et al. (2002); Siti Rahayah et al. (2008b); Snider & Styles (2003); Stobart et al. (1992) which state that the existence of DIF is associated with individual background factors such as gender, race, family's economic situation, location, facility, teacher, mother tongue and culture. Some of the factors which often become the focus of DIF researchers are gender (Cole 1997; Covic et al. 2007; Higgins 2007; Linn & Hyde 1989; Siti Rahayah et al. 2008b; Stobart & Gipps 1997) and race (Hsin Huang Li & Stout 1995; Siti Rahayah et al. 2008b; Stoneberg 2004). In a study which compares the methods for detecting DIF relating to gender, Lamprianou et al. (2002) stresses gender and language have been repeatedly mentioned as factors which might raise DIF. In a DIF study, students' demography factors are variable to determine the categories of groups being compared in the study.

The analysis of GDIF and DIF carried out on MyGSI is an effort to ensure evaluation exercise is fair for every student who undergoes it. The DIF analysis has been applicable to instruments related to the cognitive field, academic subject evaluation and applied in questionnaire research items, qualification tests, promotions, the granting of licenses and other publications (Dodeen 2004; Zieky 2002). According to Dorans & Holland (1993); Dorans & Kulick (1998); Holland & Thayer (1988); Siti Rahayah et al. (2008c), there are differences between groups who sit for evaluation in certain areas. DIF study in instrument development on evaluation education is a primary method which is aimed at identifying the differences. The tests could show no similarity or nearly identical function when it is administered to a group of similar ability students. Although the students possess similar ability, their nature and background are different (Covic et al. 2007; Higgins 2007). Based on earlier studies and items characteristics, the main cause of the DIF occurrence lies in more than one dimension (multidimensionality) measured in an item (Shealy & Stout 1993). This means that the item measures at least another dimension; apart from the principal dimension that it should measure. The unplanned dimension could exist within the type, content or method of study that is being employed.

Conclusion

Higher Learning Institution or universities are the most suitable place to build and enhance students' generic skills. The mastery of all aspects of the generic skills would facilitate in students' academic achievement. Since the university students are comprised of diverse backgrounds, the generic skills assessment needs to be carried out justly. Therefore, DIF inspection in MyGSI would classify items based on gender, types of program and race. Separation or exclusion of items that are identified by DIF would increase the reliability and validity of MyGSI instrument. In order to build students' generic skills profile, it is suggested to employ MyGSI that is free from DIF. It could also be an indicator to boost students' generic skills especially during their university years. Consequently, it is advisable to implement a larger scale research that comprises of a wider sample from all IPTA students in every state in Malaysia. This would enrich the diverse demographic backgrounds of the respondents and the research as well. In a nutshell, DIF benchmarking could also be carried out involving all of the IPTA.

References

Ackerman, T. 1992. A didactic explanation for item bias, item impact and item validity from a multidimensional perspective. *Journal of Educational Measurement* 29: 67-91.

Allan, J. & Clarke, K. 2007. Nurturing supportive learning environments in higher education through the teaching of study skills: To embed or not to embed?. *International Journal of Teaching and Learning in Higher Education* 19(1): 64-76.

American Psychological Association. 1985. *Standards for Educational and Psychological Testing*. Washington DC: American Psychological Association.

Andrich, D. 1978. Relationship between the Thurstone and Rasch approaches to item scaling. Angoff. 1993. *Applied Psychological Measurement* 3: 446-460.

Ausubel, D.P. 1978. *Educational Psychology: A Cognitive View*. (2nd ed.). New York: Holt Reinhart.

Ballantine, J. & Larres, M.P. 2007. Cooperative learning: A pedagogy to improve students' generic skills?. *Journal Education and Training* 49(2): 126-137.

Bennett, N., Dunne, E. & Carre, C. 2000. *Skill Development in Higher Education and Employment*. The Society for Research into Higher Education & Open University Press.

Berk, R.A. 1982. *Handbook of Methods for Detecting Test Bias*. Baltimore: Johns Hopkins University Press.

Biggs, J. 2003. *Teaching for Quality Learning at University.* Maidenhead: Society for Research into Higher Education & Open University Press.

Bond, T.G. & Fox, C.M. 2001. *Applying the Rasch Model: Fundamental Measurement in the Human Sciences*. New Jersey: Lawrence Erlbaum Associates Publishers London.

Bond, T.G. & Fox, C.M. 2007. *Applying the Rasch Model: Fundamental Measurement in the Human Sciences*. Second Edition. New Jersey: Lawrence Erlbaum Associates Publishers London.

Camilli, G. 1993. The case against item bias detection techniques based on internal criteria: Do item bias procedures obscure test fairness issues. In *Differential item functioning,* Holland, P.W. & Wainer, H. New Jersey: Lawrence Erlbaum Associates.

Cauffman, E. & MacIntosh, R. 2006. A Rasch differential item functioning analysis of the Massachusetts Youth Screening Instrument: Identifying race and gender differential item functioning among juvenile offenders. *Educational and Psychological Measurement* 66: 502-521.

Cole, N.S. 1981. Bias in testing. *America Psychologist* 36: 1067-1077.

Cole, N.S. 1997. The EST gender study: How females and males perform in educational settings. *EST Technical Report.*

Covic, T., Pallant, J.F., Conaghan, P.G. & Tennant, A. 2007. A longitudinal evaluation of the center for epidemiologic studies-depression scale (CES-D) in Rheumatoid Arthritis population using Rasch analysis. *Health Qual Life Outcomes* 5: 41. http://creativecommons.org/licenses/by/2.0.

Dodeen, H. 2004. Stability of differential item functioning over a single population in survery data. *The journal of Experimental* 72(3): 181-194.

Dodeen, H. & Johanson, G. 2001. The prevalence of gender DIF in survey data. *Annual Meeting Proposal of the American Educational Research Association*. April 2001. Seattle: 10-14.

Dorans, N.J. & Holland, P.W. 1993. DIF detection and description: Mantel-Haenszel and standardization. In *Differential Item Functioning,* edited by Holland, P.W. & Wainer, H. Hillsdale. N.J.: Lawrence Erlbaum Associates.

Dorans, N.J. & Kulick, E. 1998. Demonstrating the utility of the standardization approach to assessing unexpected differential item performance on the Scholastic Aptitude Test. *Journal of Educational Measurement* 23(4): 355-368.

Falk, I. & Millar, P. 2002. Implications of 'non-standard work practices' for literacy and numeracy. *ALNARC National Research Program Commonwealth of Australia.*

Fox, C.M. & Jones, J.A. 1998. Uses of rasch modeling in counseling psychology research. *Journal of Counseling Psychology* 45(1): 30-45.

Gibson, S.G. & Harvey, R.J. 2003. Gender and ethnicity based differential item functioning on the Armed Services Vocational Aptitude Battery. *Equal Opportunities International* 22: 1-15.

Hambleton, R.K. & Jones, R.W. 1994. Comparison of empirical and judgmental procedures for detecting differential item functioning. *Education Research Quarterly* 18: 21-37.

Hambleton, R.K. & Rogers, H.J. 2000. Developing an item bias review form. *University of Massachusetts at Amberst*. http://ericae.net/fit/tamu/biaspub2.htm.

Hambleton, R.K., Swaminathan, H. & Rogers, J.H. 1991. *Fundamental of Item Response Theory*. Newbury Park, CA: Sage.

Hambur, S., Rowe, K. & Luc, L.T. 2002. *Graduate Skills Assessment*. Australian Council for Educational Research. Commonwealth Department of Education Science & Training.

Higgins, G.E. 2007. Examining the Original Grasmick Scale: A Rasch model approach. *Journal Criminal Justice and Behavior* 34: 157.

Holland, P.W. & Thayer, D.T. 1988. Differential item performance and the Mantel-Haenszel procedure. In *Test Validity,* edited by Wainer, H. & Braun, H.I. New Jersey: Lawrence Erlbaum Associates.

Hsin-Huang Li & Stout, W. 1995. A new procedure for detection of crossing DIF. *Research Report*. University of Illinois at Urbana-Champaign, Illinois.

Jager, K.D. & Nassimbeni, M. 2005. Information literacy and quality assurance in South African higher education institutions. *South African Journal of Libraries and Information Science* 55: 31-38.

Kearns, P. 2001. *Review of Research: Generic Skills for the New Economy*. Adelaide: NCVER.

Lai, J.S. & Eton, D.T. 2002. Clinically Meaningful Gaps. *Rasch Measurement Transactions* 15(4): 850.

Lamprianou, I., Boyle, B. & Nelson, N. 2002. Using optimal appropriateness measurement to detect examinees mostly affected by differential item functioning in the year 2000. Key stage 2, science, national curriculum tests in England. *Educational Measurement, Psychometrics and Assessment Issue?. page number?.* University of Manchester.

Linacre, J.M. 2002. *Differential item and test functioning*. (online report) http://www.rasch.org/rmt/rmt163g.htm.

Linacre, J.M. 2007. *A User's Guide to WINSTEPS Rasch-model Computer Programs*. Chicago: MESA Press.

Linn, M.C. & Hyde, J.S. 1989. Gender, mathematics and science. *Educational Researcher* 18(8): 17-27.

Lizzio, A. & Wilson, K. 2004. First-year students' perceptions of capability. *Studies in Higher Education* 29(1): 109-128.

Lublin, J. 2003. Generic objectives and transferable skills. *Centre for Teaching and Learning*: Good Practice in Teaching and Learning.

Maller, S.J. 2001. Differential item functioning in the WISC-III: Item parameters for boys and girls in the national standardization sample. *Educational and Psychological Measurement* 61(5):793-817.

Masters, G.N. 1982. A Rasch model for partial credit scoring. *Psychometrika* 47(2): 149-174.

McNamara, T.F. 1996. *Measuring Second Language Performance*. London and New York: Addison Wesley Longman.

Nunnaly, J.C. & Bernstein, I. 1994. *Psychometrics Theory*. Ed. 3. New York: McGraw-Hill.

O'Neill, K. & McPeek, W. 1993. Item and test characteristics that are associated with differential item functioning. In *Differential Item Functioning*, edited by Wainer, H. & Braun, H.I. Hillsdale, NJ: Lawrance Erlbaum Associates, Inc.

Osterlind, S.J. 1989. *Contructing Test Items*. Boston: Kluwer Academic Publishers.

Pallant, J.F. & Tennant, A. 2007. An introduction to the Rasch measurement model: An example using the Hospital Anxiety and Depression Scale (HADS). *Br J Clin Psychol* 46: 1-18.

Rasch, G. 1960. *Probabilistic Models for Some Intelligence and Attainment Tests*. Copenhagen: Denmarks Paedagogiske Institut.

Rodiah Idris, Siti Rahayah Ariffin & Noriah Mohd Ishak. 2008a. Pemeriksaan differential item functioning (DIF) instrumen pentaksiran kemahiran generik. Seminar Psikologi dan Pembangunan Manusia: 262-273.

Rodiah Idris, Siti Rahayah Ariffin & Noriah Mohd Ishak. 2009b. A Rasch differential item functioning analysis of the Malaysian generic skills instrument. Conference of the International Journal of Arts & Sciences. 1-4 June, Austria, 1(18): 263-279.

Rodiah Idris, Siti Rahayah Ariffin & Noriah Mohd Ishak. 2009c. Application ofRasch model in validating the construct of measurement for generic skills Instrument for Higher Education (GeSIHE). Conference of the Pacific Rim Objective Measurement Symposium - PROMS 2009. 28-30 July, Hong Kong, 100-115.

Rodiah Idris, Siti Rahayah Ariffin & Noriah Mohd Ishak. 2009d. Pengaruh Kemahiran Generik dalam Kemahiran Pemikiran Kritikal, Penyelesaian Masalah dan Komunikasi Pelajar Universiti Kebangsaan Malaysia. *Malaysian Journal of Learning and Instruction* 6. (In Press).

Scheuneman, J.D. 1987. An experimental exploratory study of causes of bias in test item. *Journal of Education Measurement* 24: 97-118.

Sharifah Hapsah Syed Hasan Shahabudin. 2006. *Kerangka Kelayakan Malaysia*. Lembaga Akreditasi Negara.

Shealy, R. & Stout, W.F. 1993. A model-based standardization approach that separates true bias/DIF from group differences and detects test bias/DTF as well as item bias/DIF. *Psychometrika* 58: 159-194.

Shepard, L.A. 1982. Definition of bias. In. *Handbook of Methods for Detecting Test Bias,* Berk, R.A. Baltimore: Johns Hopkins University Press.

Sheppard, R., Han, K., Colarelli, S.M. & Dai, G. 2006. Differential item functioning by sex and race in Hogan Personality Inventory. *Assessment* 13(14): 442-453.

Siti Rahayah Ariffin, Noriah Mohd Ishak, Abdul Ghafur Ahmad, Rodiah Idris & Nur' Ashiqin Najmuddin. 2008a. Communication, leadership, and teamwork skills as core competencies among higher education students. Proceeding ASAIHL International Conference. 7-10 April, Bangkok, 149-158.

Siti Rahayah Ariffin, Noriah Mohd Ishak, Riza Atiq O.K Rahmad, Abdul Ghafur Ahmad, Rodiah Idris, Nur'Ashiqin Najmuddin. 2008b. Assessing generic skills using rasch model approach: A method for construct validity and reliability. International Conference on Education.

Siti Rahayah Ariffin, Rodiah Idris & Nur'Ashiqin Najmuddin. 2008c. Inovation using rasch model approach in measuring generic skills. International Conference on Education.

Siti Rahayah Ariffin, Noriah Mohd Ishak, Roseni Ariffin, Abdul Ghafur Ahmad & Rodiah Idris. 2008d. Evaluation approaches and challenges using structural equation model (SEM). Proceeding International Conference on the Education of Learner Diversity. 427-440.

Siti Rahayah Ariffin, Rodiah Idris & Noriah Mohd Ishak. 2008e. Profil kemahiran generik pelajar-pelajar institut pengajian tinggi: Kajian kes di Universiti Kebangsaan Malaysia (UKM). Seminar Kebangsaan Jawatankuasa Penyelaras Pendidikan Guru.

Skinner, B.F. 1971. *Beyond Freedom and Dignity.* New York: Vintage Books.

Smith, R.M. 2000. Fit analysis in latent trait measurement models. *Journal of Applied Measurement* 1(2): 199-218.

Snider, P. & Styles, I. 2003. Psychometric analysis of triandis' instrument of collectivism and individualism using modern latent trait theory. Murdoch University, Western Australia. (online report) http://www.aare.edu.au/01pap/sni01709.htm.

Stobart, G., Elwood, J. & Quinlan, M. 1992. Gender bias in examinations: how equal are the opportunities?. *British Educational Research Journal* 8(3): 261-276.

Stobart, G. & Gipps, C. 1997. *Assessment: A Teacher's Guide to the Issues.* Ed. 3. London: Hodder & Stoughton.ning (DIF) in the Spring 2003 Idaho Standards Achievement Test applying the Simultaneous Bias Test (SIBTEST) and the Mantel Haenszelchi-square test. (online report) http://www.sde..state.id.us/admin/docs/ayp/ISATDIFStudy.pdf.

Stoneberg, Jr. B.D. 2004. A Study of Gender-Based and Ethnic-Based Differential Item Functioning (DIF) in the Spring 2003 Idaho Standards Achievement Tests Applying the Simultaneous Bias Test (SIBTEST) and the Mantel-Haenszel Chi Square Test, *Internship in Measurement and Statistics:* 1-15.

Title, C.K. 1982. Test bias. In *Educational Research, Methodology and Measurement an International Handbook,* edited by Keeves, J.K. Oxford, England: Pergamon Press.

Wright, B.D. & Masters, G.N. 1982. *Rating Scale Analysis*. Chicago: MESA Press.

Zieky, M. 1993. Practical questions in the use of DIF statistics in test development. In *Differential Item Functioning,* edited by P.W. Holland & H. Wainer, 337-347, Hillsdale, N.J: Lawrence Erlbaum.

Zieky, M. 2002. Ensuring the fairness of licensing tests. Clear Exam Review 12(1): 20-26. (online report) http://www.clearhq.org/cer.htm.

Zwick, R. & Erickan, K. 1989. Analysis of differential item functioning in the NAEP history assessment. *Journal of Educational Measurement* 26: 55-66.

Siti Rahayah Ariffin
Faculty of Education
Universiti Kebangsaan Malaysia
43600 Bangi
Selangor