6 A Comparative Study of IELTS and ACCESS Test Results

Magdalena Mok and Nick Parr
Macquarie University
Tony Lee and Elaine Wylie
Griffith University,

Abstract

ACCESS is the English language examination of the Department of Immigration and Multicultural Affairs (DIMA) of the Australian Federal Government to assess migration applicants' English language skills for migration purposes. An alternative examination for the same purpose is the IELTS (General Training) module. The objective of the present study is to establish the equivalence between the measurement scales of ACCESS and IELTS (General Training). The findings would be necessary for the migration process and informative for IELTS as an international English language testing system.

The equivalence between ACCESS and IELTS can be established by administering the two tests to the same group of individuals. In such a case, only a small sample of individuals can be tested. The results from such an analysis, however, would be difficult to generalise both because the size of the sample can rarely claim representativeness and because the situation of the test administration is not typical of the actual examinations. Alternatively, actual test results from the two examination systems can be used for the establishment of the equivalence. In such a case, there is a need for links between the two sets of test results through candidates who have taken a third common test. This latter method was used in the study. The linking test chosen was the ASLPR. The IELTS (Academic) module was also included in the study to enable a complete analysis of IELTS. The analysis in the study thus included four testing systems: the IELTS (Academic and General Training) modules, ACCESS and the ASLPR.

The statistical technique used to establish the equivalence was many-facet Rasch modelling. This is an approach for equating test scales by reference to an external measurement scale independent of the tests involved. Using Rasch modelling, equivalence among the test scales can be established from actual test results of separate ACCESS and IELTS candidate groups. The results of the equating are also highly generalisable due to the statistical modelling techniques employed.

The findings from the equating exercise have enabled the identification of the scale structures of the four testing systems. The results are most interesting both in terms of the understanding gained regarding the test scales and in terms of the methodology used.

As regards the equivalence between ACCESS and IELTS (General Training), there is a large segment of match between the ACCESS and the IELTS scales. This provides sufficient basis for estimating the equivalence of the two scales.

The application of Rasch modelling in establishing equivalence among testing systems has also made a contribution to applied linguistic research.

1.0 Introduction

This is a research study jointly funded by the Department of Immigration and Multicultural Affairs (DIMA) of the Australian Federal Government and IELTS Australia to establish equivalence between English language proficiency levels specified by the ACCESS test[1] and IELTS. This final report has been prepared for both the Steering Committee of the ACCESS project and IELTS Australia. Findings are generally applicable to both testing systems. Implications and conclusions, however, may be different and will be reported in separate reports to DIMA and IELTS Australia.

ACCESS is designed specifically for the assessment of English language ability required for migration to Australia,(see Appendix 6.1 for level details) whereas IELTS is for individuals seeking entrance into higher education in English medium institutions (principally in the UK and Australia). It can, however, be used as an alternative to ACCESS for migration to Australia. Recently, the New Zealand government has adopted IELTS as their assessment instrument for screening migration applications regarding English language ability.

The question naturally arises as to whether and to what extent equivalence can be established between ACCESS and IELTS results. In fact, one of the concerns of DIMA regarding research into ACCESS was in establishing such equivalence. From the point of view of ACCESS, the equivalence would have implication for test administration and, by consequence, the migration process itself. As regard IELTS, the relative degrees of match and mismatch with other English language assessment systems would be invaluable for concurrent and even construct validity.

As the relevance of the study and its findings may be different for DIMA and IELTS Australia, separate reports have been prepared, also to do justice to the separate funding for the project from both bodies. The research team would like to express its appreciation to both DIMA and IELTS Australia for funding and support of the study. In particular, various IELTS Test Centres have been supportive in providing IELTS results for the study; so too have also been several AMES centres in providing ratings on the ASLPR. The study could not have been possible without such widespread support.

2.0 The Research Question

The research question to be addressed in the study is the following:

> "What would be the equivalence between the levels assigned to candidates taking the ACCESS test to those assigned to candidates taking the IELTS (General Training Module)?"

That is a simple enough question. The complexity is in obtaining comparable data and in specifying a common standard of comparison among tests operating with different scales.

At the outset, it may be opportune to emphasise that, as two distinct English language testing systems, the equivalence between ACCESS and IELTS cannot be computed directly, but has to

[1] The ACCESS test is the English language proficiency examination of the Australian Federal Government for migration purposes.

be estimated. The two testing systems follow different test design principles and systems of test scores computation. The matching of the two testing systems, therefore, can be made by reference to a standardising scale distinct from the scales in the two testing systems.

The definition of the standardising scale, though, is not a simple matter. It has to be such that it can establish equivalence between tests (IELTS and ACCESS) with different measurement scales. There is also an issue relating to all behavioural measurement scales. These are generally structured as scales with numerical levels. For example, the IELTS scale has nine levels (see Appendix 6.2 for further detail):

Level 1	non user
2	intermittent user
3	extremely limited user
4	limited user
5	modest user
6	competent user
7	good user
8	very good user
9	expert user.

The impression created by using numerical levels is that the scale levels are of equal distance. That is rarely, if ever, the case. For instance, it cannot easily be assumed that a candidate at IELTS *Level 6* is twice as able as another at *Level 3*. The structure of the scales used in any educational or language tests, thus, has to be estimated and cannot be assumed to have levels with equal intervals.

3.0 The Sample

It would appear that to obtain data for matching ACCESS and IELTS, it would be necessary to select a sample of individuals to whom the two tests would be administered. That was, however, neither feasible nor desirable for the current study. It was not feasible because both tests are secured public examinations and could not be administered outside the official context for test security reasons. In addition, the complexity and the cost in administering both tests also made sample testing not feasible. It was also undesirable to run sample testing. A sample of the required size and representativeness would be difficult to collect. Furthermore, sample testing could not easily claim to replicate the actual test situation. A sample design not relying on administering both tests to a sample of candidates was thus the only option possible.

An alternative approach in obtaining the required data would be to use actual test results. The establishment of equivalence would be primarily a statistical process without requiring that all the tests involved be taken by all candidates in the matching. It would be sufficient for some candidates to have taken any two of the tests. This is known as common-person equating.

Various techniques exist to equate measurement scales from separate candidate groups. The particular sample selection would be dependent on the estimation model adopted. In the current study, the statistical approach used was item response modelling (IRM). In that context, the equivalence between different assessment scales from separate candidate groups could be established via links in the data provided by common-person equating. The links could be individuals who had done any two of the tests involved. The number of candidates in the linking group was not critical within item response modelling. The linking candidates served to

provide a common reference for the external measurement scale used in the establishment of test scale equivalence. This is known as the *common frame of reference* within IRM.

The sample for the current study was composed of past ACCESS and IELTS candidates and individuals assessed using the Australian Second Language Proficiency Ratings (ASLPR). The links in the data were provided by individuals having been assessed by ACCESS or IELTS and the ASLPR. There were usually few individuals who had been assessed by both ACCESS and IELTS. The ASLPR was, thus, the measurement where the links in the data were found (see Appendix 6.3 for ratings).

The equivalence to be established was primarily between ACCESS and IELTS (General Training Module) as the two alternative assessment instruments for migration to Australia. As far as IELTS was concerned, the inclusion of the Academic Module in the matching would be highly desirable to provide a full picture of the equivalence. In addition, the inclusion of the Academic Module would also provide an answer to requests from various individuals about using IELTS (Academic Module) also for migration purposes.

The sample entered into the analysis consisted, thus, of sub-samples of candidates assessed by the two IELTS modules, ACCESS and the ASLPR. The total number of candidates in the sample was 2,093 assessed on the four macro-skills of Listening, Reading, Writing and Speaking, with 502 in ACCESS, 759 in the ASLPR, 477 in IELTS (Academic) and 355 in IELTS (General Training). The number of individuals assessed by the ASLPR and one of the other tests was as follows: 6 in ACCESS, 25 in IELTS (Academic) and 1 in IELTS (General Training). The distribution of the sample is summarised in Table 1:

The ACCESS sub-sample was from all ACCESS test centres during the period February to December 1995. The ASLPR sub-sample was mostly from ratings collected from 1990 to 1996 and stored in the ASLPR database of the Language Testing and Curriculum Centre (LTACC) at Griffith University. The IELTS sub-samples were test results from Australian IELTS test centres in 1995.

	ACCESS	ASLPR	IELTS A	IELTS GT
		6	6	
ACCESS	502			
ADELAIDE				52
AU100		1	1	
AU140		1	1	
BANGKOK			1	
BRISBANE			171	13
BTR		1		
CALL		732		
CANBERRA		4	281	32
INDONESIA			1	
MELBOURNE				101
MOSCOW		1		1
PERTH				57
QUT		3		
REGENCY TAFE		1		
SYDNEY				99
THAILAND			1	
UQ		9	12	
UTS			1	
VIETNAM			1	
Total	**502**	**759**	**477**	**355**

Table 1 The sample

The linking data were collected from Australian IELTS test centres and a number of Australian migrant English centres. Students in those centres were contacted to identify those with both the ASLPR ratings and either IELTS or ACCESS test scores. There were few cases of individuals with the ASLPR rating, and ACCESS or IELTS (General Training). The small number of individuals in those two sub-samples, though, did not constitute serious problems for the analysis as the links were used only to establish a common frame of reference and would not affect the calibration of the scales.

While both the ACCESS and the IELTS sub-samples were from a number of sources and represented a variety of individuals, the ASLPR sub-sample was predominantly from one centre. That might affect the estimation of the ASLPR scale, even if the estimation of other scales would not be affected. The interpretation of the ASLPR scale structure, thus, would have to be cautious.

4.0 Item Response Modelling (IRM)

As discussed above, the analysis model to be adopted in this study should be one that is capable to match the four tests with only partial data overlap through the link in the ASLPR ratings. The measurement model chosen has to map the four separate scales onto a common scale without requiring that all candidates be tested on all four tests. In addition, the distinctness of the four scales have to be maintained while they are being matched onto a common external scale. The calibration of the scales, as described above, needs also to be independent of the sample upon which the scale calibration is carried out. This is to ensure general applicability of the results of the comparison. The use of IRM ensures that can be achieved.

Item Response Modelling (IRM) is a statistical process whereby behavioural measurement data can be given meaning consistent with the behavioural assessment context. By this it is meant that the raw scores obtained in any behavioural measurement, eg a language test, are not taken at their face value. The raw scores have to be re-interpreted statistically by taking into consideration factors that may affect them. The particular factors to be taken into account vary from situation to situation. The resulting IRM is labelled as an n-facet IRM. There are a number of variations of IRM. The one that is most commonly employed in behavioural sciences is the Rasch model. Many-facet Rasch models were used in this study.

By using many-facet Rasch models, the four scales could be matched onto a common external scale because there were candidates with an ASLPR rating together with one of the other three tests. The ASLPR ratings provided the links across the other three tests. Because of the calibration procedure in Rasch modelling such partial overlap in the data was sufficient to enable calibration. This was achieved by a mathematical transformation of the raw scores called the logit transformation. What a logit transformation does is to transform the raw score into the probability. In the simplest two-facet model consisting of candidate ability and test item difficulty a candidate's ability is expressed in terms of the probability of that candidate answering right an item of zero difficulty; the difficulty level of an item, in turn, is expressed in terms of the probability of the item being answerable by a candidate with zero ability. The logit transformation results in a common external scale for the calibration of all facets in the model. Consequently, whatever the scales of the tests involved, the calibration is referenced to a common scale; so are other facets in the model. If there are links in the data, all the elements in the model would be calibrated on a single underlying logit scale, providing an unambiguous interpretation for the results.

IRM provides a solution to two issues in all behavioural measurements. The first is the probabilistic nature of all behavioural measures, and the second is the unequal spacing of all rating scales. The fact that numerical scores are assigned to behavioural measurement data would give the impression that the scores represent a certain level of ability in a direct way. That is not the case. The assignment of a certain level on a rating scale is best conceptualised as the placement of an individual on the rating scale on the basis of the raw results from the test. That can be achieved only in probabilistic terms. The situation just described is true for both subjectively rated and objectively marked tests. In the latter case, the raw score (number of items correct) cannot be taken as the direct measure of the ability tested but rather as an indicator of a certain level of the ability measured via a selection of test items. The number of test items answered right in any test bears no one-to-one relationship with the ability measured, but is the basis upon which the level of ability mastered by the candidate can be estimated. In an intuitive way, the probabilistic nature of behavioural measurement is what we generally

perceived as what happens when we give a global and impressionistic assessment to human behaviour. The following observation by Linacre (1993)[2] may be opportune:

> ".. the ultimate goal of the judging process ... is not to determine some 'true' rating for an examinee on each item, on which ideal judges would agree, but rather to estimate the examinee's latent ability level, of which each judge's rating is a manifestation." (p.41)

In a similar way, rating scales are usually expressed in terms of scales with levels in equally spaced intervals. That gives the impression that the levels in the scales are also equally spaced. That, again, is not the case. In the strict sense, rating scale levels are qualitative labels of distinct human behaviour with ordered performance. An individual scoring a *four* on a rating scale is not necessarily twice as able as an individual scoring a *two*. The progression from one level on a rating scale to another, therefore, cannot be taken to be of equal distance. In what ways, particular rating scales are actually paced, has to be investigated empirically.

IRM estimates the ability of a person and the difficulty of a test item in terms of a mathematical model representing the probability of success. The most important result of the estimation is a model of measurement that would represent the data to the highest degree of probability. The model is prescriptive in that it is the best case scenario for the data that can be identified. Naturally, the fit between the model and the data is never perfect. The degree of fit, then, has to be estimated and the relative patterns of fit diagnosed. *Identification, estimation* and *diagnosis* are, in fact, the three basic steps in statistical modelling. The model is thus based on the data without being restricted by it, and indicates the underlying patterns within the data. The processes of estimation and diagnosis of the model would enable an assessment of the explanatory power of the model.

The probabilistic scale resulting from the IRM (the logit scale) constitutes a standard measure, like the meter or the thermometer. It can be used to measure all aspects of the behavioural assessment situation in a uniform way. In addition, the probabilistic scale can encompass a complex assessment situation. For example, rater severity, instruments using different measurement scales or examinee background can all be included in an item response model with estimation and diagnosis for all of the characteristics (facets). The levels in the rating scales in the model are also estimated, resulting in scale levels referenced to the logit scale.

5.0 The Item Response Model in the Study

In the model for test equivalence in the study, the four macro-skills were taken as four measurements (items) of a single underlying scale. The four tests involved (IELTS - Academic and General Training, ACCESS and the ASLPR) were considered four different scales measuring the same set of common macro-skills. That captured the general understanding of the assessment situation. In referring to results in the four tests in question, general statements like *Peter is 5 in Reading, 4 in Listening and Speaking and 5.5 in Writing on IELTS* are used. That assumes a single underlying scale in IELTS.

The item response model adopted had the four macro-skills as four ratings across the four tests. Each test, though, was estimated using its own scale. By doing so, the scale structure of the

[2] Linacre J.M. 1993 *Many-Facet Rasch Measurement*. MESA, Chicago.

four tests was preserved even if they were calibrated onto a single underlying logit scale. That was how the equivalence among the scales could be established.

Three facets were included in the model: *candidate*, *the four tests* and *the four macro-skills*. Each facet was involved in the model estimation without influencing one another, a standard estimation procedure in IRM. That was very aptly summed up by Linacre (1993: 41) as follows

> "it is possible to obtain ... an estimate of the ability of each examinee, freed from the level of severity of the particular judges who happened to rate the performance and also freed from the difficulties of the items and the arbitrary manner in which the categories of the rating scale have been defined."

The above is applicable to all facets in any item response model and is defined as the local *independence* requirement of IRM.

The four English language proficiency scales estimated were the best estimates of the scale structure within the scope of the particular model used. They represented the operationalisation of the overall scale structure of the tests. The logit scale against which the four individual scales were calibrated served as the standard for the establishment of equivalence.

The overall scale structures were the combined estimates of the four macro-skills. There were expected to be differences between the overall scale and the scales for the four macro-skills individually. It would be rather unusual to assume that the macro-skills were equally demanding in any language test. Estimates of deviations of the macro-skills from the overall scale could be estimated and were used to derive the scales for the macro-skills.

The many-facet Rasch model to be estimated can be expressed as follows:

$$\log\left(\frac{P_{nijk}}{P_{nijk-1}}\right) = B_n - D_i - C_j - F_k$$

where

P_{nijk} is the probability of candidate n receiving on macro-skill i in test j a level of k.
P_{nijk-1} is the probability of candidate n receiving on macro-skill i in test j a level of $k-1$.
B_n is the ability of candidate n.
D_i is the difficulty of macro-skill i.
C_j is the difficulty of test j.
F_k is the difficulty of the step up from level $k-1$ to level k.

As a prescriptive model, the parameters estimated bore different degrees of fit with the data. A lack of fit by itself did not necessarily invalidate either the model or the data. It indicated only that the data were found to be not fully in line with the model. Different treatments can be applied depending on the situation. In the case of estimating candidate performance, those candidates misfitting the model should be further looked at to determine the possible nature of the misfit. It may be the case that some of those misfitting candidates are candidates with special personal or group characteristics. In the case of item calibration, misfitting items are generally items that need to be re-designed or eliminated from the test. The purpose of model

fitting is not to eliminate the misfits but to be able to identify misfits and to further investigate the situation. In some situations estimating a different model may even be necessary.

Within the present study, the degrees of fit among the scales would indicate the extent of agreement among them, as the main objective of the study is to evaluate the equivalence between ACCESS and IELTS (General Training) in terms of the levels critical for the migration process. It would be sufficient if the relevant levels in both the ACCESS and the IELTS scales could be matched rather than having the two scales matching in all levels. It is not expected that the four scales under examination will be fully matched.

The model was fitted using the many-facet Rasch package FACETS with the following model specifications:

```
Title = ACCESS, IELTS and ASLPR Comparison
Data file = ielts96a.DSP
Output file = ielts96a.out

; Data specification
Facets = 3
Non-centered =  1
Positive =  1
Labels =
 1,Candidate (elements = 2067)
 2,Test (elements = 4)
 3,Macro_Skill (elements = 4)
Model =?,#B,?B,R90,1

; Output description
General statistics (point bi-serial) = yes
Unexpected observations reported if standardized residual >= 3

; Convergence control
Convergence = 5, .01
Iterations (maximum) = 120
Xtreme scores adjusted by = .3, .5  ;(estimation, bias)

; Data Summary
Total lines in data file = 2093
Responses matched to model: ?,#B,?B,R90 = 8372
    Total non-blank responses found = 8372
Responses with unspecified elements = 0
 Responses not matched to any model = 0
Valid responses used for estimation = 8372
```

6.0 The Frame of Reference

The equivalence between the ACCESS and the IELTS scales is established by matching the two scales along a common measurement dimension shared by them. The measurement dimension is estimated using the Rasch model in 5.0 above. The aim is not to match the scales in all levels. It is sufficient that a relevant dimension common to the scales being matched is estimated. In terms of the current study, the critical levels in the ACCESS and the IELTS scales, where decisions on migration are dependent, would be the segments of the scales that need to be fitted with the model. This constitutes the frame of reference for the current study. This is important for the interpretation of the results and for the evaluation of the study.

It is certainly possible to use different configurations of the data entered into the model building or to gradually refine the model to be estimated to achieve fully matched scales. That was, however, beyond the scope of the current study.

7.0 Results

7.1 General Fit of the Model

The following were the unexpected responses from the model fitted:

	Macro-Skills				Grand
Test	**Listening**	**Reading**	**Speaking**	**Writing**	**Total**
ACCESS	2	2	2	1	**7**
ASLPR	29	0	0	9	**38**
IELTS_A	2	0	1	1	**4**
IELTS_G	2	1	5	2	**10**
Grand Total	**35**	**3**	**8**	**13**	**59**

Table 2 Unexpected responses

The total number of unexpected responses of 59 should be considered rather low indicating a general good fit of the model. Among the four tests, the largest number of unexpected responses are found in the ASLPR with Listening having the highest number, 29, followed by Writing, 9. Other tests have very few unexpected responses, which can be taken as random. The general fit of the model is as below:

```
Count of measurable responses = 8284
Count of independently estimable parameters = 2090
Data-to-model global fit:
log-likelihood chi-square: 22386.3 d.f.: 6194 significance: .001
residual chi-square: 8385.1  d.f.: 6194  significance: .001
```

Both the log-likelihood and the residual chi-square statistics are significant at the 0.1% level. The model has reasonable overall fit. It should be pointed out, though, that the overall fit is of little consequence for the results of a study. It only indicates a general picture. It is the relative fit in specific parameters that would have bearing on the estimation.

7.2 Overall Model Calibration

The result of the model calibration is summarised in Table 3. The leftmost column is the logit scale. The scale is expressed in terms of equal-paced steps representing various degrees of probability with zero representing a fifty-fifty probability. It is not pertinent here to detail the degrees of probability associated with each level of the scale. It is sufficient that the scale provides a basis for comparing the facets in the model. The *Candidate* facet is positively oriented, with the upper end of the scale representing higher levels of candidate ability; the other facets are negatively oriented with upper end of the scale representing higher degrees of difficulty.

By reading across the columns, all the facets can be compared in a uniform way. For example, the peak of the *Test* facet has the ASLPR as the test where it is difficult to achieve a high level rating. The IELTS modules are placed in the middle and ACCESS the lowest on the logit scale. The calibration of the four test scales will be discussed in detail with examination of the values in the four rightmost columns in Table 3.

Logit	+Candidate	-Test	-Macro-Skill	IELTS A	IELTS G	ASLPR	ACCESS
7	.			(9)	(9)	(5)	(6)
	.						
6							

	.						
5	.						
	.					4+	
				---	---		
4							
	.				8.5	---	
	.			8.5			
				---	---		----
3	.			8	8		
	*.			---	---		
	.				7.5		5
	*.			7.5	---	4	
2	**			---	7		
	.			7	---		----
	**.			---	6.5		
	**.			6.5	---	---	
1	*.			---			
	****.	ASLPR		6	6	3+	4
	*.		Writing	---	---		
	****		Listening	5.5		---	
0	***.	IELTS_G			5.5		
	***.	IELTS_A	Reading Speaking	---	---	3	----
	****.	ACCESS		5			
	*****.			---	5	---	
-1	*********.			4.5	---	2+	3
	****.						
	***.			---	4.5	---	----
	*.			4	---	2	
-2	.				4	1+	
	.			---		1	2
	.			3.5	---	1-	
	.				3.5	---	
-3	.			---	---		----
	.			3	3		
	.			---	---		
					2		
-4				2			
	.			---	---		
-5	.			(1)	(1)	(0)	(1)

Logit	* = 33	-Test	-Macro_Skill	IELTS	ASLPR	ACCESS

Table 3 The Model Calibration

7.3 The Overall Test Scale Structure

The test scale calibration in Table 3 is based on the four macro-skills as repeated measurements using a common scale. Each test, however, has its own specific scale. The overall scale, thus, represents only the general scales of the four tests and represents the understanding that there is a common scale being referred to for the macro-skill assessment in each of the tests. After calibration, the levels in the four scales can be matched by reference to the logit scale.

It can be seen that there is a rather close match between the scales of the two IELTS modules. The levels in the two scales match one another quite well. ACCESS *Level 4* matches roughly IELTS *5.5* to *6.5*, and ACCESS *5* with IELTS *7* to *8*. However, this is only a general match.

As a prescriptive model, the calibration in Table 3 is expected to have varying degrees of fit with the scales operationalised in the data. The relative degrees of fit have to be estimated.

The diagnosis of the calibration for the four test scales is as follows in Tables 4a to 4d. (The operationalisation of the IELTS scale in the two modules has to be treated as two different observed measurements. Their correspondence needs to be investigated empirically.)

The first column is the *Scale* of the respective tests. The *Logit* column contains the estimates of the difficulty of the levels in the scales.

The *Outfit* column indicates the precision of the logits. The outfits are indices of consistency of ratings.[3] The expected value of the outfits is 1 and the conventional limits for the outfits to be acceptable are between 0.7 to 1.3, representing two standard errors above and below the estimates. This is the confidence interval comprising 95% of the cases in the model. Outfits below 0.7 actually represent higher consistency in the ratings than expected. They do not present problems for the model, but indicate that the estimates are restricted to the data on hand and cannot be generalised. The outfit values are related to the conformity of the levels in the scale to the model and have, thus, to do with the validity of the levels to the model estimated. In the current study, outfits lower than 0.7 were not considered generalisable for establishing equivalences.

The *-0.5* column contains the logits for the thresholds between levels in the scale. The levels are presented as discrete; the ratings, however, are continuous. The levels in a scale are, in reality, bands. It is thus necessary to estimate the points where one level ends and the next begins. The logits in the column refers to the thresholds below the levels associated. Thus, the column label of -0.5 being the lower threshold of a level, the value is not calculated for the lowest level.

The next three columns contain standard error statistics. The first includes the standard errors (*S.E.*) estimated. The *-2 S.E.* and *+2 S.E.* columns are the lower and the upper limits of the confidence intervals for the scale levels. They indicate the limits of variation of ratings.

[3] Consistency in ratings refers to the pattern in the ratings where the persons with higher ability receiving higher ratings and vice versa. Within Rasch models ratings are expected to have a certain degree of inconsistency. Ratings with too high a degree of consistency are thus not expected.

The + /- *2 S.E* values have to be between the *-0.5* thresholds of the scale level associated and the level above to be considered sufficiently precise. In other words, the confidence interval needs to be within the band width of the scale level. The three *S.E.* columns, thus, make up the reliability statistics of the scale level.

The *Band* column reports the band width for the levels in the scale. The bands are in logits and reflect the span of ability a particular level covers. The two extreme levels are, in theory, indefinite in width and are thus without a band width. The band widths represent the degree of English language ability required to pass from one level on the scale to another. The wider the width the greater the degree of ability is required.

The statistics taken together provide a good idea of the structure of the rating scales.

The scale structure of IELTS (Academic Module) in Table 4a has the following features:

The scale covers a range from -4.80 to 4.62 logits. The span of the scale is rather balanced on both the positive and the negative poles. The centre of the scale is at *Level 5.5* with a logit of 0.19, not too distant from the theoretical centre of 0. The scale is, thus, balanced.

Level	Logit	Outfit	-0.5	S.E.	-2 S.E.	+2 S.E.	Band
9	4.62	3.2	4.13				
8.5	3.60	6.0	3.24	0.10	3.40	3.80	0.89
8	2.93	2.4	2.65	0.16	**2.61**	**3.25**	0.59
7.5	2.37	2.1	2.11	0.07	2.23	2.51	0.54
7	1.86	1.4	1.60	0.07	1.72	2.00	0.51
6.5	1.34	1.0	1.07	0.03	1.28	1.40	0.53
6	0.79	0.6	0.49	0.02	0.75	0.83	0.58
5.5	0.19	0.7	-0.13	0.01	0.17	0.21	0.62
5	-0.46	1.0	-0.78	0.06	-0.58	-0.34	0.65
4.5	-1.11	1.6	-1.45	0.04	-1.19	-1.03	0.67
4	-1.81	1.7	-2.18	0.13	-2.07	-1.55	0.73
3.5	-2.54	2.0	-2.88	0.10	-2.74	-2.34	0.70
3	-3.20	2.2	-3.53	0.18	**-3.56**	**-2.84**	0.65
2	-3.89	1.8	-4.38	0.05	-3.99	-3.79	0.85
1	-4.80	2.3					

Table 4a IELTS (Academic Module) scale

From the Outfits in Table 4a it can be seen that only *Levels 5*, *5.5* and *6.5* in the IELTS (Academic Module) fit the model and can be used for estimating equivalences across the four scales.

The standard errors that fall outside the confidence interval (higher than the threshold of the next higher level and lower than the scale level) are bold-faced. *Level 8* is wider than the band width at both the upper and the lower limits. Ratings at that level vary too widely to be reliable. It should be observed that the upper threshold is exceeded by only 0.01 logit. Likewise, *Level 3* has also the confidence interval wider than the band width in both ends.

The bands are evenly spaced except the two outer-most levels with a band width over **0.80**. The average band width is **0.65** logits.

The scale structure of IELTS (General Training Module) shown in Table 4b covers a range from -5.05 to 4.95 logits. Here too, the scale has a balanced span on both poles. The effective centre of the scale is again *Level 5.5* with a logit of 0.07, very close to the theoretical centre.

Level	Logit	Outfit	-0.5	S.E.	-2 S.E.	+2 S.E.	Band
9	4.95	2.7	4.36				
8.5	3.74	2.2	3.35	0.13	3.48	4.00	1.01
8	3.05	1.3	2.79	0.14	**2.77**	3.33	0.56
7.5	2.56	1.3	2.34	0.06	2.44	2.68	0.45
7	2.11	1.0	1.87	0.08	1.95	2.27	0.47
6.5	1.58	1.1	1.24	0.04	1.50	1.66	0.63
6	0.84	0.9	0.44	0.05	0.74	0.94	0.80
5.5	0.07	1.0	-0.30	0.01	0.05	0.09	0.74
5	-0.69	0.8	-1.08	0.04	-0.77	-0.61	0.78
4.5	-1.45	1.0	-1.80	0.04	-1.53	-1.37	0.72
4	-2.12	1.6	-2.41	0.14	-2.40	-1.84	0.61
3.5	-2.66	3.8	-2.90	0.10	-2.86	-2.46	0.49
3	-3.15	2.7	-3.43	0.17	**-3.49**	**-2.81**	0.53
2	-3.81	5.1	-4.44	0.14	-4.09	-3.53	1.01
1	-5.05	2.2					

Table 4b IELTS (General Training Module) Scale

The outfit values between *Levels 4.5 to 8* are within the acceptable range. There is, therefore, a large middle segment of the scale fitting the model.

The lower bound of *Level 8* is 0.02 (**2.77** to 2.79) logit below the lower threshold of that level. The deviation is, however, very small. *Level 3*, on the other hand, has a confidence interval wider than the band width.

The bands are similar to the Academic scale with evenly spaced intervals. The two outer-most levels (*Levels 2 and 8.5*) have a band of **1.01**. The average band width is **0.68** logit.

The ACCESS scale structure in Table 4c has a span from -3.83 to 3.94 logits, balanced on both poles. The effective centre of the scale is between *Levels 3* and *4* at -0.885 logit, half way between -1.00 to 0.77 logits and is close to the theoretical centre.

Level	Logit	Outfit	-0.5	S.E.	-2 S.E.	+2 S.E.	Band
6	3.94	1.5	3.28				
5	2.46	0.7	1.73	0.03	2.40	2.52	1.55
4	0.77	0.7	-0.26	0.03	0.71	0.83	1.99
3	-1.00	1.0	-1.61	0.06	-1.12	-0.88	1.35
2	-2.24	1.3	-3.09	0.09	-2.42	-2.06	1.48
1	-3.83	2.4					

Table 4c The ACCESS Scale

The scales levels have all acceptable confidence intervals and are thus reliable.

The outfit values between *Levels 2 to 5* are all within the acceptable limits, leaving only the two extreme levels not fitting the model. The bands are evenly spaced, with *Level 4* being the widest level with a span of **1.99** logits. The average band width is **1.59** logits.

The ASLPR scale structure in Table 4d[4] spans from -2.84 to 6.31 logits.

Level	Logit	Outfit	-0.5	S.E.	-2 S.E.	+2 S.E.	Band
5	6.31	1.1	5.55				
4+	4.57	2.8	3.64	0.20	4.17	4.97	1.91
4	2.33	3.3	1.28	0.21	1.91	2.75	2.36
3+	0.65	1.8	0.16	0.09	0.47	0.83	1.12
3	-0.28	1.0	-0.71	0.07	-0.42	-0.14	0.87
2+	-1.10	0.3	-1.42	0.03	-1.16	-1.04	0.71
2	-1.64	0.4	-1.8	0.02	-1.68	-1.60	0.38
1+	-1.92	0.2	-2.04	0.02	-1.96	-1.88	0.24
1	-2.15	0.3	-2.27	0.01	-2.17	-2.13	0.23
1-	-2.41	0.6	-2.63	0.01	-2.43	-2.39	0.36
0	-2.84	1.7					

Table 4d The ASLPR Scale

Unlike the other scales, there is an imbalance on both poles. The effective centre of the scale is at *Level 2+*, which has a logit of -1.10. The centre of the scale is thus considerably below the zero-logit theoretical centre. The scale as a whole, therefore, has a very narrow lower segment with a very wide upper segment.

[4] The full ASLPR scale has a **0+** level. This is not found in this analysis as there have not been any ratings at the **0+** level.

The outfit values have only *Levels 3* and *5* within the acceptable range. The model estimated, thus, fits only at these two levels in the ASLPR scale. *Levels 1-* to *2+* are all with over-fit values (values below 0.7).

The confidence intervals of the levels are all narrower than the band widths. The scale levels are, thus, all sufficiently reliable.

The band widths are divided into three segments. *Levels 1-* to *2* are very narrow (ranging from **0.24** to **0.38**); *Levels 2+* and *3* have band widths of **0.71** and **0.87** respectively; *Levels 3+* to *4+* have band widths between **1.12** to **2.36**. The scale has, thus, uneven band widths. In such a case, it may not be advisable to consider the mean band width.

The scale structures described above are represented graphically in Figures 1a to 1c.

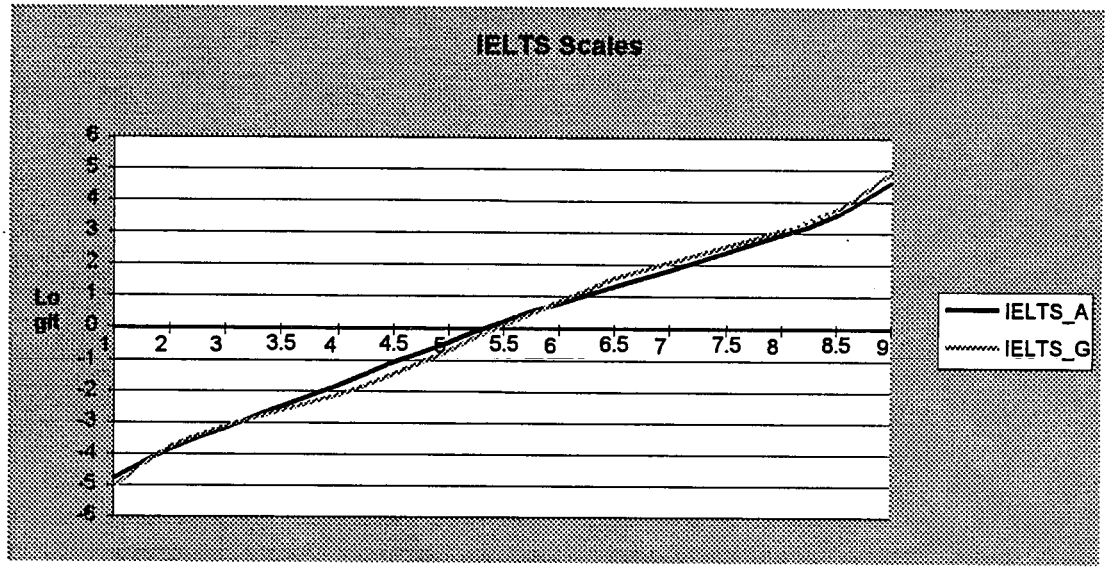

Figure 1a *The IELTS Scales*

The IELTS scales are placed in the same graph because they are operationalisations of the same underlying scale, and have been found to have similar structures. Putting the two graphs together would thus provide a good comparison of them. Indeed, the two graphs are nearly identical in Figure 1a above. The IELTS scale appears to be very close to a linear structure with a slightly inverted S shape.

The graphs cut across the **0** logit point at about *Level 5.5* and the upper and the lower segments are well balanced.

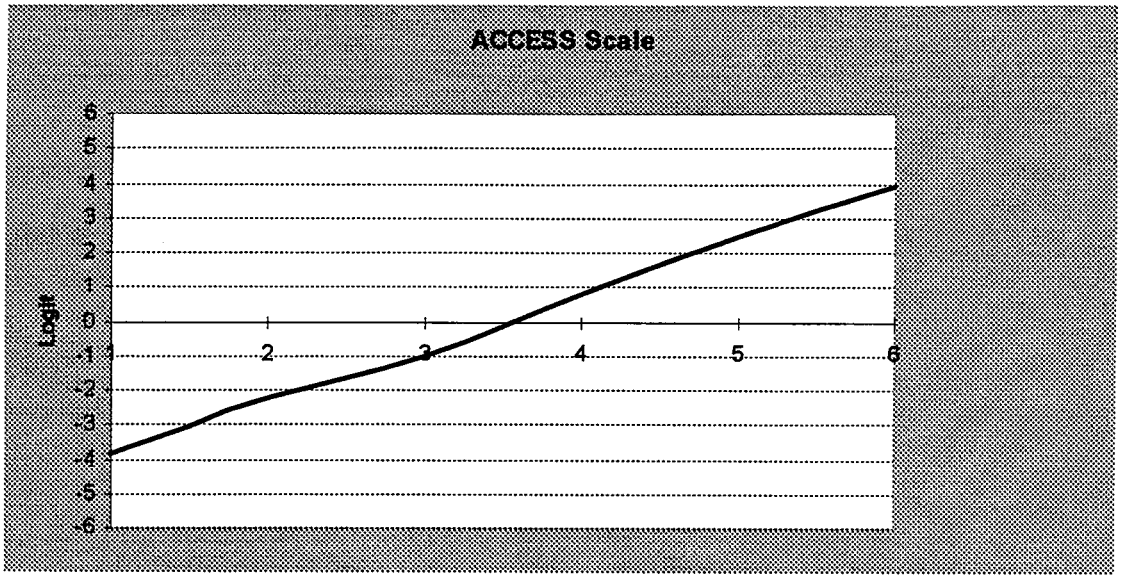

Figure 1b *The ACCESS Scales*

The ACCESS scale has also a near-linear structure. The graph crosses the **0** logit point between *Levels 3* and *4*. The upper and the lower segments are also well balanced.

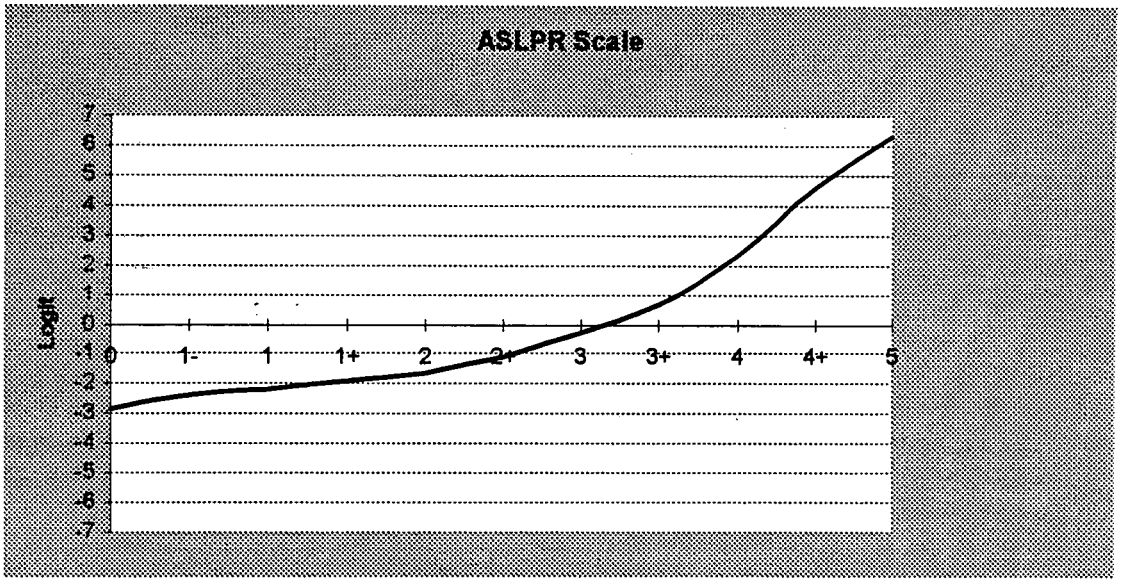

Figure 1c *The ASLPR Scales*

The general match of the ACCESS and the IELTS scales can be deduced by inspecting the corresponding logit levels of the scales. ACCESS *Level 4* is at about +1 logit, which corresponds to *Level 6* on the IELTS scale. ACCESS *Level 5* is at about **+2.5** logit, which matches the point on the IELTS scale between *Levels 7.5* to *8*. These results are very close to those reported in 7.2 above.

The ASLPR scale displays a non-linear structure. The graph is rather flat to the left of zero logit (*Levels 0* to *3*), but takes up a very steep climb after crossing the zero logit point (*Levels 4* to *5*).

7.4 The Fit of the Scale Structures

The scale structures estimated in 7.3 are the modelled scales for the four tests in question and represent the best estimates of the generalisable scale structures. Naturally, the modelled structures have varying degrees of fit with the data on hand. The model and the degrees of fit have to be interpreted within the frame of reference of the particular modelling process and should not be given any absolute interpretation. Any lack of fit only indicates a lack of conformity between the data and the model and does not necessarily imply a poor scale.

As indicated in Section 7 above, it is sufficient to have the critical levels in the two scales concerned match along a common measurement dimension estimated with the Rasch model. In terms of the results from the FACETS analysis, those levels in the scales with outfit values between 0.7 to 1.3 will be used for establishing the equivalence between the ACCESS and the IELTS scales.

The levels in the four scales with acceptable outfit values are as follows:

IELTS (Academic Module):	*Levels 5, 5.5 and 6.5.*
IELTS (General Training Module):	*Levels 4.5* to *8.*
ACCESS :	*Levels 2* to *5.*
ASLPR:	*Levels 3* and *5.*

7.5 Macro-Skill Specific Match

The analysis of the overall scale structures and the specification of equivalence described so far refer to the overall scales, which have been estimated using the ratings of macro-skills in each of the tests as measurements of the overall scales. Such an approach is justified as the scales are used in assessing all the macro-skills. It is, however, understood that the four macro-skills cannot be assumed to be equally difficult in any tests. This is so either from the point of view of first or second language acquisition, or from test design and implementation, or from candidate performance in any language tests. It is thus expected that there are deviations from the overall scale structures specified in the macro-skills. Such deviations can be estimated within FACETS, using *bias* analysis. This is generally known as residual analysis, which has the general form of

```
Observed Value = Estimated Value + Residual Value
```

Once the Residual values for the micro-skills are estimated, the scales for the individual macro-skills can be specified using the formula above. Table 5 reports the residuals (biases) in the macro-skills for the four tests.

It should be observed that the outfit values are all within the acceptable range. In the *Bias+ Logit* column, a negative value indicates that the particular macro-skill concerned is easier than the overall scale while a positive value indicates the opposite.

From the results in Table 5, Listening in IELTS (Academic), for example, is more than half a logit (-0.64) easier than the overall scale, IELTS (General Training) is about one quarter

(-0.37) logit easier, the ASLPR is 0.22 logit harder and ACCESS 0.61 logit harder than the overall scale.

The scales for the individual macro-skills are derived by the following formula:

```
Macro Skill Logit = Overall Logit + Macro Skill Bias Logit
```

Test	Macro-Skill	Bias+ Logit	Outfit MnSq
IELTS A	Listening	-0.64	0.8
IELTS A	Reading	0.44	0.7
IELTS A	Writing	0.12	0.9
IELTS A	Speaking	0.09	1.0
IELTS G	Listening	-0.37	1.0
IELTS G	Reading	0.34	1.1
IELTS G	Writing	-0.06	0.9
IELTS G	Speaking	0.09	1.1
ASLPR	Listening	0.22	1.2
ASLPR	Reading	-0.24	0.6
ASLPR	Writing	0.00	1.1
ASLPR	Speaking	-0.13	0.5
ACCESS	Listening	0.61	0.8
ACCESS	Reading	-0.64	0.9
ACCESS	Writing	-0.26	0.8
ACCESS	Speaking	0.25	0.9

Table 5 Residuals

The results of the scales associated with the macro-skills are reported in Tables 6a to 6d on the following pages, one for each of the tests in question.

	Listening		Reading		Writing		Speaking	
Level	Logit	-0.5	Logit	-0.5	Logit	-0.5	Logit	-0.5
9	3.98	3.49	5.06	4.57	4.74	4.25	4.71	4.22
8.5	2.96	2.60	4.04	3.68	3.72	3.36	3.69	3.33
8	2.29	2.01	3.37	3.09	3.05	2.77	3.02	2.74
7.5	1.73	1.47	2.81	2.55	2.49	2.23	2.46	2.20
7	1.22	0.96	2.30	2.04	1.98	1.72	1.95	1.69
6.5	0.70	0.43	1.78	1.51	1.46	1.19	1.43	1.16
6	0.15	-0.15	1.23	0.93	0.91	0.61	0.88	0.58
5.5	-0.45	-0.77	0.63	0.31	0.31	-0.01	0.28	-0.04
5	-1.10	-1.42	-0.02	-0.34	-0.34	-0.66	-0.37	-0.69
4.5	-1.75	-2.09	-0.67	-1.01	-0.99	-1.33	-1.02	-1.36
4	-2.45	-2.82	-1.37	-1.74	-1.69	-2.06	-1.72	-2.09
3.5	-3.18	-3.52	-2.10	-2.44	-2.42	-2.76	-2.45	-2.79
3	-3.84	-4.17	-2.76	-3.09	-3.08	-3.41	-3.11	-3.44
2	-4.53	-5.02	-3.45	-3.94	-3.77	-4.26	-3.80	-4.29
1	-5.44		-4.36		-4.68		-4.71	

Table 6a *The IELTS (Academic Module) Macro-Skill Scales*

	Listening		Reading		Writing		Speaking	
Level	Logit	-0.5	Logit	-0.5	Logit	-0.5	Logit	-0.5
9	4.58	3.99	5.29	4.70	4.89	4.30	5.04	4.45
8.5	3.37	2.98	4.08	3.69	3.68	3.29	3.83	3.44
8	2.68	2.42	3.39	3.13	2.99	2.73	3.14	2.88
7.5	2.19	1.97	2.90	2.68	2.50	2.28	2.65	2.43
7	1.74	1.50	2.45	2.21	2.05	1.81	2.20	1.96
6.5	1.21	0.87	1.92	1.58	1.52	1.18	1.67	1.33
6	0.47	0.07	1.18	0.78	0.78	0.38	0.93	0.53
5.5	-0.30	-0.67	0.41	0.04	0.01	-0.36	0.16	-0.21
5	-1.06	-1.45	-0.35	-0.74	-0.75	-1.14	-0.60	-0.99
4.5	-1.82	-2.17	-1.11	-1.46	-1.51	-1.86	-1.36	-1.71
4	-2.49	-2.78	-1.78	-2.07	-2.18	-2.47	-2.03	-2.32
3.5	-3.03	-3.27	-2.32	-2.56	-2.72	-2.96	-2.57	-2.81
3	-3.52	-3.80	-2.81	-3.09	-3.21	-3.49	-3.06	-3.34
2	-4.18	-4.81	-3.47	-4.10	-3.87	-4.50	-3.72	-4.35
1	-5.42		-4.71		-5.11		-4.96	

Table 6b *IELTS (General Training Module) Macro-Skill Scales*

Level	Listening		Reading		Writing		Speaking	
	Logit	-0.5	Logit	-0.5	Logit	-0.5	Logit	-0.5
6	4.55	3.89	3.30	2.64	3.68	3.02	4.19	3.53
5	3.07	2.34	1.82	1.09	2.20	1.47	2.71	1.98
4	1.38	0.35	0.13	-0.90	0.51	-0.52	1.02	-0.01
3	-0.39	-1.00	-1.64	-2.25	-1.26	-1.87	-0.75	-1.36
2	-1.63	-2.48	-2.88	-3.73	-2.50	-3.35	-1.99	-2.84
1	-3.22		-4.47		-4.09		-3.58	

Table 6c The ACCESS Macro-Skill Scales

Level	Listening		Reading		Writing		Speaking	
	Logit	-0.5	Logit	-0.5	Logit	-0.5	Logit	-0.5
5	6.53	5.77	6.07	5.31	6.31	5.55	6.18	5.42
4+	4.79	3.86	4.33	3.40	4.57	3.64	4.44	3.51
4	2.55	1.50	2.09	1.04	2.33	1.28	2.20	1.15
3+	0.87	0.38	0.41	-0.08	0.65	0.16	0.52	0.03
3	-0.06	-0.49	-0.52	-0.95	-0.28	-0.71	-0.41	-0.84
2+	-0.88	-1.20	-1.34	-1.66	-1.10	-1.42	-1.23	-1.55
2	-1.42	-1.58	-1.88	-2.04	-1.64	-1.80	-1.77	-1.93
1+	-1.70	-1.82	-2.16	-2.28	-1.92	-2.04	-2.05	-2.17
1	-1.93	-2.05	-2.39	-2.51	-2.15	-2.27	-2.28	-2.40
1-	-2.19	-2.41	-2.65	-2.87	-2.41	-2.63	-2.54	-2.76
0	-2.62		-3.08		-2.84		-2.97	

Table 6d The ASLPR Macro-Skill Scales

The deviation scales of the macro-skills are summarised in Table 7.

The levels with good fit are marked with a dark background.

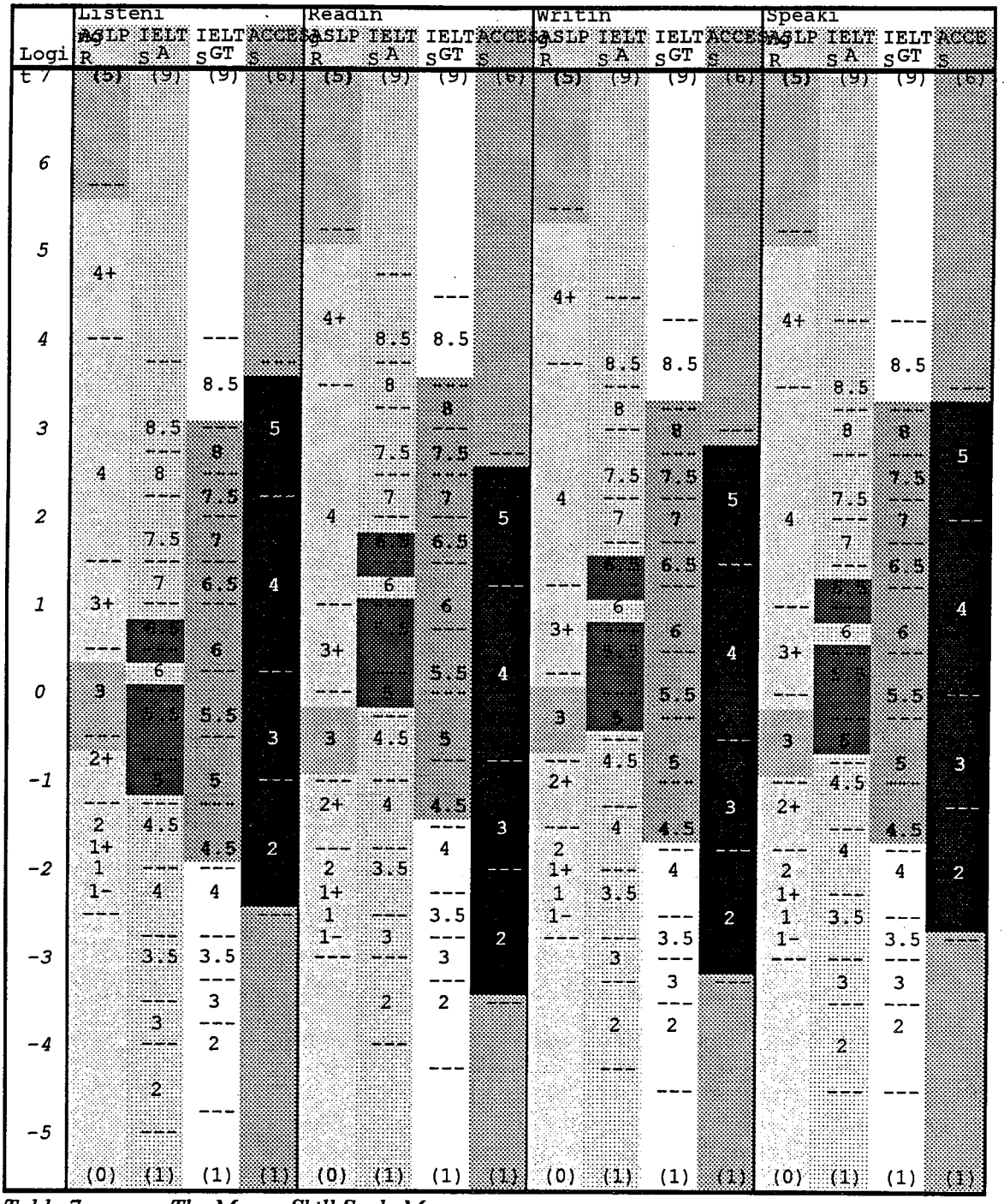

Table 7 The Macro-Skill Scale Map

Among the four tests, ASLPR is the most difficult test, the IELTS modules are next in difficulty and ACCESS is the easiest of the four tests.

In the ASLPR, the four macro-skills are of similar difficulty levels with Listening being the most difficult albeit in a small degree. The two IELTS modules have rather similar patterns of difficulty levels. Listening is the easiest of the four macro-skills. In ACCESS, on the other hand, Listening is the most difficult of the four macro-skills. Such discrepancies in the patterns of difficulty levels among the macro-skills in the four tests justify the use of different matching

rules for each of the macro-skills rather than the use of a single set of rules of equating across all macro-skills.

8.0 Equivalence between IELTS and ACCESS

Detailed specification of equivalence of the macro-skills between ACCESS and IELTS (General Training Module) is reported below in Table 8:

ACCESS Level	Top	Bottom	IELTS Upper Limit	Width	From -0.5	Lower Limit	Width	From -0.5
Listening			***Listening***					
5	3.89	2.34	*8*	(0.56)	*0.91	*7.5*	(0.45)	0.37
4	2.34	0.35	*7.5*	(0.45)	0.37	*6*	(0.80)	0.28
3	0.35	-1.00	*6*	(0.80)	0.28	*5*	(0.78)	0.45
2	-1.00	-2.48	*5*	(0.78)	0.45	*4.5*	(0.72)	-0.31
Reading			***Reading***					
5	2.64	1.09	*7*	(0.47)	0.43	*6*	(0.80)	0.31
4	1.09	-0.90	*6*	(0.80)	0.31	*4.5*	(0.72)	0.56
3	-0.90	-2.25	*4.5*	(0.72)	0.56	*4.5*	(0.72)	0.31
Writing			***Writing***					
5	3.02	1.47	*8*	(0.56)	0.29	*6.5*	(0.63)	0.29
4	1.47	-0.52	*6.5*	(0.63)	0.29	*5*	(0.78)	0.62
3	-0.52	-1.87	*5*	(0.78)	0.62	*4.5*	(0.72)	-0.01
Speaking			***Speaking***					
5	3.53	1.98	*8*	(0.56)	*0.09	*7*	(0.47)	0.02
4	1.98	-0.01	*7*	(0.47)	0.02	*5.5*	(0.74)	0.20
3	-0.01	-1.36	*5.5*	(0.74)	0.20	*4.5*	(0.72)	0.35
2	-1.36	-2.84	*4.5*	(0.72)	0.35	*4.5*	(0.72)	-1.13

Table 8 Equivalence between ACCESS and IELTS

The left side of Table 8 contains the ACCESS levels and their upper (the **Top** column) and lower (the **Bottom** column) thresholds. The right-side IELTS section contains the upper limit of the IELTS levels matching the ACCESS **Top** with the width of the IELTS level concerned (the **Width** column) and the distance from the lower threshold (the **From -0.5** column). This is followed by the same pieces of information for the lower limit of the IELTS level corresponding to the **Bottom** of the ACCESS level. For example, *Level 5* in ACCESS Reading covers the range from 1.09 (**Bottom**) to 2.64 (**Top**) logits This range matches at the **Bottom** with IELTS *Level 6* at 0.31 logit from the lower threshold and at the **Top** with IELTS *Level 7* at 0.43 logit from the lower threshold. By subtracting the **Width** (0.47) from the **From -0.5** value (0.43) for the upper limit, it can be deduced that the match at the upper limit is at

0.04 logit below the **Top** of *Level 7* in IELTS. A precise point of match between the ACCESS and the IELTS levels can thus be determined.

The same computations can be made for the **Bottom** of the matched levels. The values in the **from -0.5** column for the upper threshold with an asterisk (Listening and Speaking) indicate that the **Top** of the ACCESS levels is actually above the IELTS level (*Level 8*). *Level 8.5* has not been matched because that level does not have an acceptable outfit value in the overall calibration and has not been included in the estimation of equivalences. Similar treatment is applied for ACCESS levels below the acceptable range of outfit values at the **Bottom**. In such cases (Listening, Writing and Speaking) the values in the **From -0.5** column are reported in the negative.

Two ACCESS levels (*Levels 5* and *4*) are critical for ACCESS test administration. These are discussed in details below. ACCESS *5* in Listening covers a range from 2.34 to 3.89 logits, which matches a range from above IELTS *8* to the middle of *7.5*. The actual upper limit of ACCESS *5* is 0.91 logit above the ceiling of IELTS *8* and at 0.10 logit below the upper limit of IELTS *8.5*, which is 90% of a full IELTS *8.5* (band width 1.01). The lower bound of ACCESS *5* is at 0.37 logit from the lower boundary of that level, representing 18% of that level. ACCESS *4* spans from 0.35 to 2.34 logits. The upper boundary reaches 82% of IELTS *7.5*, its lower boundary 65% of IELTS *6*.

In Reading, ACCESS *5* spans from 1.09 logit to 2.64 logit. The upper bound is 91% of IELTS *7* and lower bound 61.25% of IELTS *6*. ACCESS *4* matches at its upper bound 38.75% of IELTS *6* and at its lower bound 22% of IELTS *4.5*.

In Writing, ACCESS *5* covers a range from 1.47 to 3.02 logits, with its upper limit reaching 52% of IELTS *8* and its lower limit 54% of IELTS *6.5*. ACCESS *4* ranges from -0.52 to 1.47 logits. The upper bound reaches 46% of IELTS *6.5* and the lower bound 21% of IELTS *5*.

In Speaking, ACCESS *5* has a range from 1.98 to 3.53 logits. The upper limit sits at 9% of IELTS *8.5* (consequently only the full IELTS *8* is used as the top of the well fitted IELTS scale). The lower limit is at 95.75% of IELTS *7*. ACCESS *4* covers -0.01 to 1.98 logits with upper limit standing at 4.25% of IELTS *7* and lower limit standing at 73% of IELTS *5.5*.

The equivalence between ACCESS and IELTS (General Training Module) has thus been established.

9.0 Discussion

9.1 The Sample

The sampling in the current study is opportunistic rather than rigorously designed. The only exception is, possibly, the sub-sample of ACCESS candidates, where all test centres have been sampled using a simple random design. However, even that sub-sample is far from scientifically drawn. Strictly scientific sample design principles have not been thought necessary because the study is not concerned with establishing estimates of the sub-populations of candidates. Resource limits have also made rigorous sampling not feasible. Difficulties in getting data was another factor making fully scientific sampling impracticable.

The sample that resulted is adequate for the study as the sub-sample for the IELTS (General Training Module) includes most Australian test centres. The sub-sample for IELTS (Academic Module) is predominantly from Canberra (281) and Brisbane (183), and the sub-sample for the ASLPR is mainly from one centre. The results from the study certainly lessen the uncertainty about the sample. The large outfit values in both the IELTS (Academic Module) and the ASLPR scale pose problems because it is not clear whether the large outfit values are due to masked centre effects because of the small number of test centres involved or a true measurement construct effect. The large outfit values, though, do not invalidate the calibration, but are ambiguous because of possible centre effects. The study as a whole should still be considered successful because of adequate sampling in both ACCESS and the IELTS (General Training Module).

9.2 The Overall Calibration

The process to establish the equivalence between the ACCESS and the IELTS scales used in this study is about the most rigorous to date for language test scale matching. Equivalence between the scales of the two testing systems have almost exclusively relied on academic opinions and anecdotal evidence. The current study is the first serious attempt in using quantitative analysis techniques on actual test data to specify that equivalence. As a measurement scale external to the facets in a calibration model, the logit scale operates in a fashion not dissimilar to a ruler or a thermometer. One of the advantages of using an external measurement scale is to enable test scale matching without having to rely on common-person sampling for the tests systems concerned. An external measurement scale for test matching enables a stable criterion of comparison independent of the things being matched.

The overall calibration includes all facets in the model, which can be referenced to the logit scale. By virtue of the linking elements in the data, the elements in the *Candidate* facet, for example, can be compared even if they are taken from different testing systems. The possibility is, thus, open for evaluating ACCESS and IELTS candidates on the same scale using a well fitted model as reference. ACCESS 1995 Research Project No.1 has demonstrated the development of such a highly generalised test administration model for ACCESS. The estimation for a similar model for ACCESS and IELTS is highly desirable.

9.3 The Fitted Segments of the Scales

The overall scale structures for the four tests (Tables 4a to 4d) show small standard errors. In those levels where the second standard errors do extend beyond the threshold for the levels, the deviations are always less than 0.1 logit. The calibration has thus a rather high degree of reliability.

The fitted segments in the four scales include large sections of the ACCESS and the IELTS (General Training Module) scales. This helps to achieve the objective of the current study. It also provides a basis for an evaluation of the validity of the calibration. From the well fitted segments across the four scales, it can be deduced that the construct measured by the model relates to a common measurable dimension between ACCESS and IELTS (General Training Module). It may not be simple to articulate in definite terms what constitutes that common dimension as it can include a complex set of characteristics. It can, however, be conjectured that the common dimension may relate to English language ability distinct from the ability to handle academic English. This is deducible because of the larger overlap between ACCESS and IELTS (General Training) than between either one with IELTS (Academic) in terms of

model fit. That is certainly an interesting finding for the discussion of core and specific-purpose language ability in applied linguistics.

In the establishment of equivalence, outfit values lower than 0.7 (the over-fit) have not been included. As indicated in 7.3, over-fit values do not, in themselves, invalidate the model. If the over-fit values are considered, IELTS (Academic Module) has levels *5* to *6.5* accepted by the model, including *Level 6* with an outfit value of 0.6. The ASLPR scale in particular has *Levels 1-* to *2+* included in addition to *Levels 3* and *5*, which have acceptable outfit values. In such a case only *Levels 0, 4* and *4+* in the ASLPR scale would not be accepted by the model. The level *0+* is also not included because of lack of any ratings at that level.

The large number of levels in the ASLPR scale with over-fit are very probably the result of a test centre effect. If there were more test centres included in the calibration, there may be more levels with acceptable outfit values. A similar deduction can be made regarding IELTS (Academic Module).

9.4 The Residual Analysis

The results from the residual analysis associated with the macro-skills call for caution in a popular practice to make simplistic matching of testing systems across all macro-skills. There does not seem to be a simple way to match test results by reference to the general scales of the testing systems. In the first place, it is justified and commonsensical to assume that candidates have different levels of ability across the macro-skills. Second, the operationalisation of the macro-skills in the test items/tasks does not automatically confirm whatever language ability levels specified in the test specifications. Finally, measurement variability and error in test administration would inevitably distort any conceptual equivalence in the test designs. Because of the above reasons, equivalence among testing systems has to be established empirically using actual test data.

10.0 Conclusion and Recommendations

The match between ACCESS and IELTS, that has been estimated in this study, has provided useful information on the current rules for matching the two testing systems within the migration process for Australia. In general terms, the rules seem to be a sufficient general guide. Refinements in those rules are certainly in order, based on the findings from the study.

The study has provided a demonstration of the application of IRM in matching ACCESS and IELTS results. It has provided important insight into the methodology of test matching. Rigorous test matching studies are still rather rare in applied linguistics. The study has, thus, implications for language testing.

The research team wishes to make the following recommendations:

- that a comprehensive model for test results reporting comprising of both ACCESS and IELTS (General Training) candidates, similar to the one developed in ACCESS 1995 Research Project No.1, be developed;
- that a weighting system be devised for the macro-skills to be applied to the current rules for ACCESS and IELTS equivalence.