言語テストにおける受験者の応答の多様性の分析
An Investigation of Response Pattern Differences in a Language Proficiency Test

中村　優治 NAKAMURA, Yuji
● 慶應義塾大学
Keio University

Keywords 特異項目機能，解答パターンの分析，個々のテスト受験者
Differential item functioning, response pattern investigation, individual test takers

ABSTRACT

This paper attempts to investigate how thirty items that function differently for three groups of students (Business Majors, Economics Majors, and Law Majors) can be identified through the differential item functioning analysis (DIF).

Traditionally, differences in performance between groups of students have been examined with the intention of identifying whether or not one group of students has been disadvantaged relatively to another on traits of the test. However, through this DIF analysis it can be pointed out that groups of students perform differently on an item. The emphasis can be shifted from the tests to the items.

Through the investigation of unexpected response patterns of the respondents' observed scores, we can find significantly different interactions among the students for some of the items, and eventually this result can be used for the improvement of the test.

本稿では，differential item functioning という手法を用いて，学生の専攻とその学習背景がテストの項目の内容とどのように関連しているのか，また個々の学生の解答パターンの中で何か特徴的なものがあるのかについて考察する．伝統的には，受験者全体がテスト全体にどのように反応したか，個々の項目にどのように反応したかという分析が一般的であるが，DIFの視点から項目を分析することで，受験者の特性と項目の特性との関連にも目を向けてテスト結果が検討できることになる．まだ，初歩的な分析の段階ではあるが，継続的な研究が望まれる分野である．

1 Purpose

The investigation of differential item functioning (DIF) has attracted a great deal of attention from test developers because DIF items pose a considerable threat to the validity of tests. DIF items should be understood to refer to test items that show significantly different interactions among different groups of students for those item. In particular, it is crucial to detect DIF items in language proficiency tests in which test takers with diverse backgrounds are involved. Also, the examination of each individual student's response patterns gives an important information to teachers who give feedback (test scores as well as advice) to their students.

Traditionally, differences in performance between groups of students have been examined with the intention of identifying whether or not one group of students has been disadvantaged relative to another on traits of the test. However, it is pointed out that groups of students perform differently on an item. The emphasis can be shifted from the tests to the items.

This paper provides an example of how thirty items that function differently for three groups of students (Business Majors, Economics Majors, and Law Majors) can be identified through the Rasch Model and the paper also shows individual test takers' unexpected response patterns.

Note 1: Differential Item Functioning (cf. —The fact that the relative difficulty of an item is dependent on some characteristic of the group to which it has been administered, such as first language or gender.) in *Multilingual glossary of language testing terms*: Studies in Language Testing 6)

Note 2: Bias (cf.—A test or item can be considered to be biased if one particular section of the candidate population is advantaged or disadvantaged by some feature of the test or item which is not relevant to what is being measured. Sources of bias may be connected with gender, age, culture, etc.) in *Multilingual glossary of language testing terms*: Studies in Language Testing 6)

2 Research Design and Method

Subjects: 62 (N=all)

Majors: Econ=Economics (n=33)
Bus=Business Administration (n=21)
Law=Law (n=8)

Test Material
30 items Dichotomous data, unscored. (1-20 had 4 categories, 21-25 had 4 categories. 25-30 had 5 categories).
Items 1-20 Grammar
Items 21-30 Cloze Test

3 Procedure of the Analysis

3.1 Examination of Validity and Reliability of the test along with the graphs of Item Characteristic Curve, Differential Item Functioning and Person-Item Location Distribution.

First, we will look at some of the basic psychometric properties of the test, that is, its validity (a question of whether it fits the Rasch model) and its reliability (the person separation index). Second, we will look at the question of whether the three student groups perform differently on each of the items (the Differential Item Functioning: DIF). This is important because if groups are performing very differently

at the item level, then those items are said to be biased, i.e. they will favour one or more groups above others. This is not good if one wants to select students on the basis of test results. However, it can be used to diagnose special areas of strength or weakness that particular groups may have in order to target teaching to them specifically.

Note: The samples here are very small – and too small to make firm judgements about the test items. The data is used, rather, illustratively, to show how the model can provide information for gathering information about student groups and about the items that could help in decision-making.

So, first, the validity and reliability of the test will be considered for all three groups combined.

It is best to consider Fit (validity – whether the test items form a single dimension or continuum) first. If a set of items (a scale) is not valid, then the reliability indices are meaningless. Therefore, the Chi square and logit residual tables will be shown first to check the following:

1) Chi square values will be examined to look for where the Chi square increases suddenly.
2) The logit residual tests of fit will be investigated. If the value is <- 3.00 or > 3.00, (in the present case, it would be preferable to take the range of -2.00 and 2.00 to be much severer) then we should look at this item further, especially if Chi square is also significantly large.
3) Along with the fit statistics, the location order will be examined This information shows the items in order of location from easiest to most difficult and indicates the wording of the easiest and most difficult for interest. It shows which skills the students in general found easy and which they had trouble with. It can also be used as a check on the face validity of the items if they are ordered as we might expect.

Next, the reliability will be examined through the Separation index which is the equivalent of Cronbach's Alpha.

3. 2 Examination of unexpected response patterns of individual students

Through the investigation of unexpected response patterns of misfitting items in terms of the students' observed scores, we might find a significant interaction between the students and the items, and eventually this result can be used for the improvement of the test and the feedback for the students.

4 Results and Discussion

4.1 Validity

Let us conduct the tests of fit of the items to the Rasch model through Chisquare and Fit Residual tests of fit. The tables below show that the item fit looks good, but this may partly be due to the small number of students in the sample. A larger sample may show more misfit in some items. At present, more data are needed to make this analysis more certain.

Table1. Chiquare

Table of Chiquare

Item	ChiSq	DF	Prob
4	0.139	2	0.932949
25	0.156	2	0.9249
15	0.205	2	0.902751
5	0.216	2	0.897678
10	0.236	2	0.888856
18	0.283	2	0.867916
23	0.318	2	0.852939
24	0.965	2	0.617235
14	1.033	2	0.596642
28	1.088	2	0.580526
27	1.367	2	0.504811
19	1.377	2	0.502252
30	1.806	2	0.405421
26	1.938	2	0.379514
11	1.942	2	0.37872
2	2.069	2	0.355323
6	2.332	2	0.311541
3	2.416	2	0.298862
13	2.947	2	0.229114
1	3.014	2	0.221526
21	3.081	2	0.214264
20	3.126	2	0.20952
17	3.17	2	0.204998
22	3.32	2	0.190103
12	4.413	2	0.110062
16	5.193	2	0.074536
8	5.548	2	0.062408
29	5.638	2	0.059654
9	6.034	2	0.048955
7	7.547	2	0.022971

Table of Fit Statistics

item	FitStat
8	−1.069
16	−0.905
24	−0.471
26	−0.44
10	−0.353
22	−0.321
12	−0.319
6	−0.299
11	−0.257
20	−0.245
9	−0.235
7	−0.118
27	0.156
18	0.171
3	0.334
30	0.42
4	0.518
25	0.565
14	0.643
5	0.663
19	0.746
15	0.807
23	0.87
1	0.969
21	1.2
29	1.204
28	1.369
13	1.91
2	1.918
17	1.967

Next, let us look at the location order which shows the items in order of location from easiest to most difficult and indicate the wording of the easiest and most difficult for interest. It shows which skills the students in general found easy and which they had trouble with. It can also be used as a check on the face validity of the items if they are ordered as we might expect.

This table indicates that item 11 is the most difficult followed by item 30 and item 27, while item 20 is the easiest followed by item 3 and item 10.

Table2. Location Order

easy

Item	Location
20	−2.209
3	−1.375
10	−1.331
22	−1.143
16	−0.99
29	−0.794
9	−0.673
1	−0.666
19	−0.256
6	−0.249
25	−0.247
12	−0.085
14	−0.038
15	0.068
4	0.094
2	0.101
8	0.107
18	0.29
28	0.294
7	0.301
13	0.337
24	0.371
26	0.435
17	0.578
5	0.583
23	0.63
21	1.104
27	1.427
30	1.468
11	1.869

difficult

4.2 Reliability

Reliability can be investigated by checking the person separation index which is the Rasch equivalent of Cronbach's alpha.

Separation Index	0.332
Cronbach Alpha	0.354

A Person Separation Index value of 0.332 is very low. Thus, the test does not discriminating very much among students The distribution graph below shows this is mainly because the test is rather too easy for most of the students in this sample. However, this test may be appropriate as a test of mastery of the particular concepts it tests. Thus discriminating among students may not be its purpose.

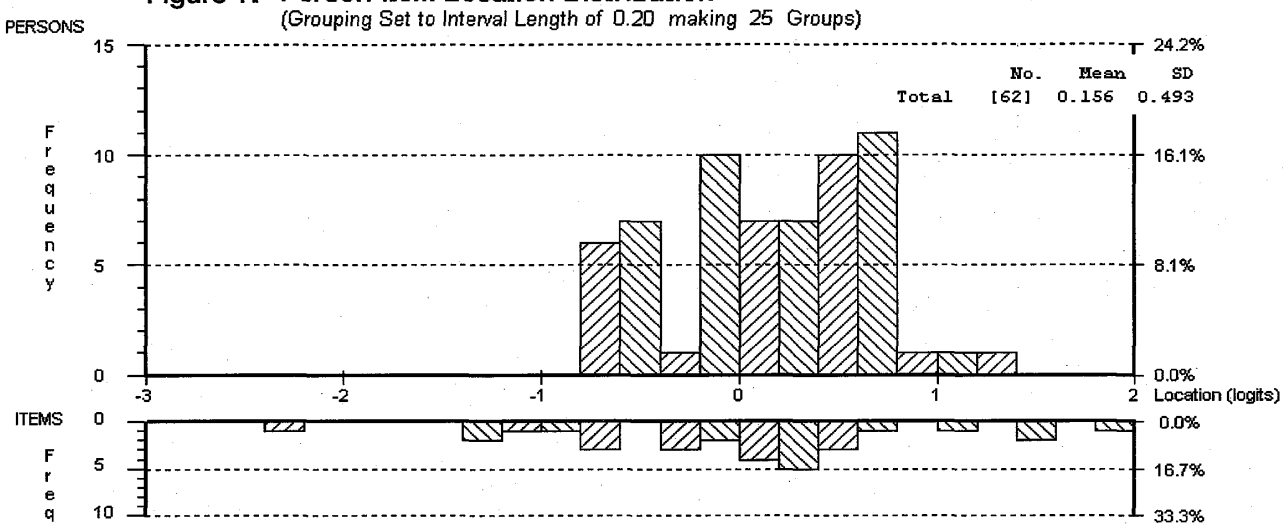
4.3 Discussion of Item Characteristic Curve (ICC)

For the validity examination we can also check the ICC curve for the item to see the actual deviation of the groups of scores (that is, the obtained values on the ICCs picture). This information is about the validity of the items, i.e. the fit to the model.

Generally speaking, there are three basic categories as follows: Category One (too flat) that shows low discrimination (ie high chi square and very positive log residual); Category Two (normal, natural) that shows good fit to the model; and Category Three (too good) that shows discrimination that is too good (ie chi square and high negative residual). The high positive discrimination is also not so desirable as it indicates some people have specialist knowledge (not part of what we are trying to assess) which advantages them with respect to others being tested.

In the present research, let us pick out some sample items in three types: Type One (items which seem to have little discriminating power) ,

Figure 1. Person-Item Location Distribution



Type Two (items which show good fit to the model) and Type Three (items which seem to have rather strong discriminating power).

Type One (item 2, item 13, item 17, item 20, item 21, item 11, item 20, item 27, item 30). The obtained characteristic curve (represented by the dots) across all locations of the students (i.e. across the operating range of the variable – i.e. horizontal axis) is too flat, relative to the expected curve (represented by the smooth line) to discriminate among people grouped into three categories (dots in the graph) on the basis of their totals scores on the whole test. i.e. the three groups have increasing mean total scores on the test as a whole.

Note: These three groups are different from the three discipline groups (law, economics and business) – they are combined groups. The three groups (three dots) are formed according to the three total score groups. In other words, the groups (dots) have increasing mean total scores on the test as a whole. In the graph the first dot from the left (n= 24), the second dot (n= 24) and the third dot (n=14).

Figure 2.
ICC Graph of Type One (e.g. item 2)

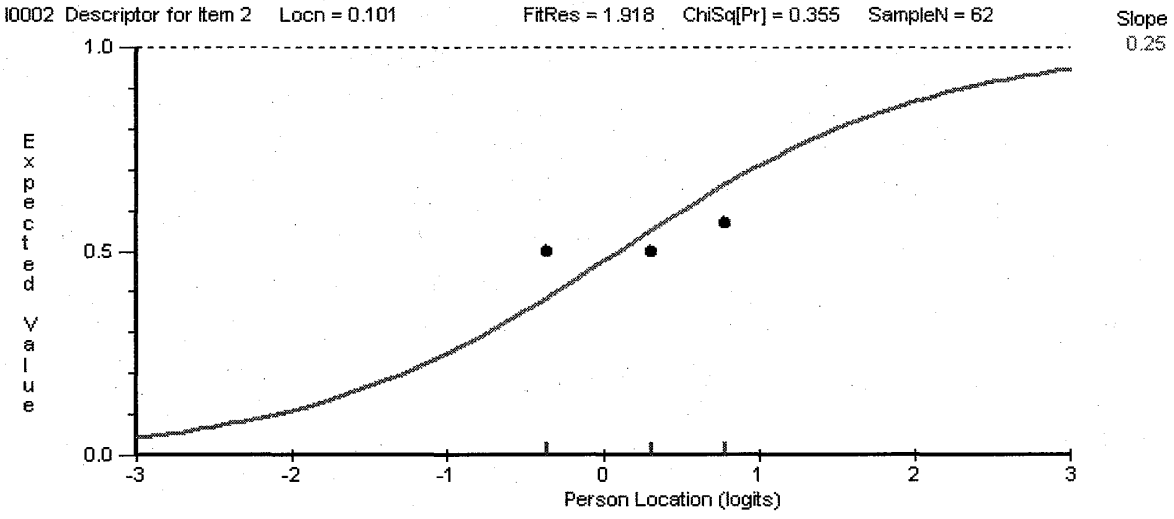
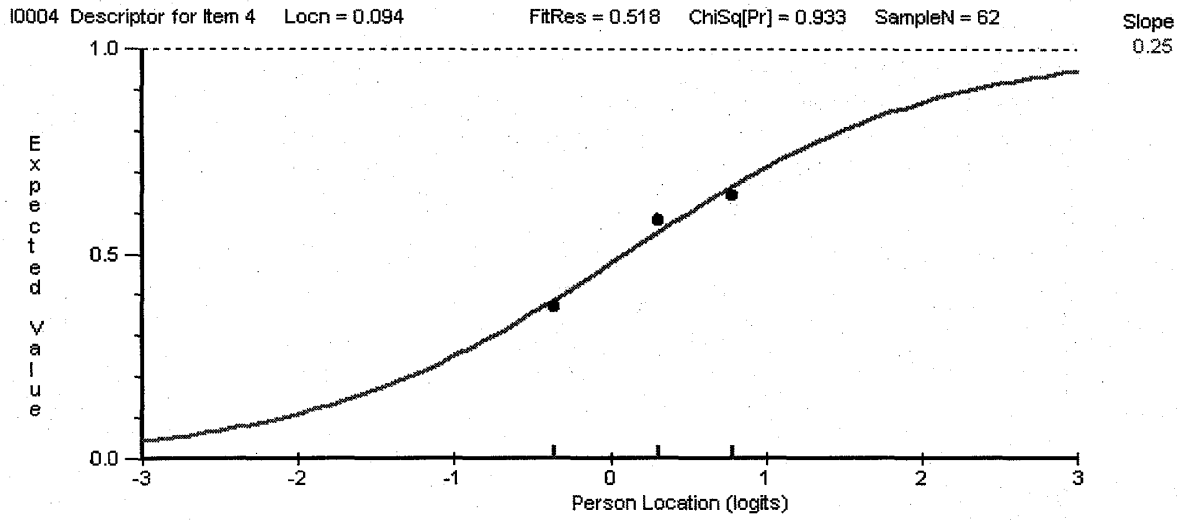


Figure 3. Type Two (item 4, item 5, item 15, item 23, item 23)
The obtained curve shows good fit to the model.
ICC Graph of Type Two (e.g. item 4)



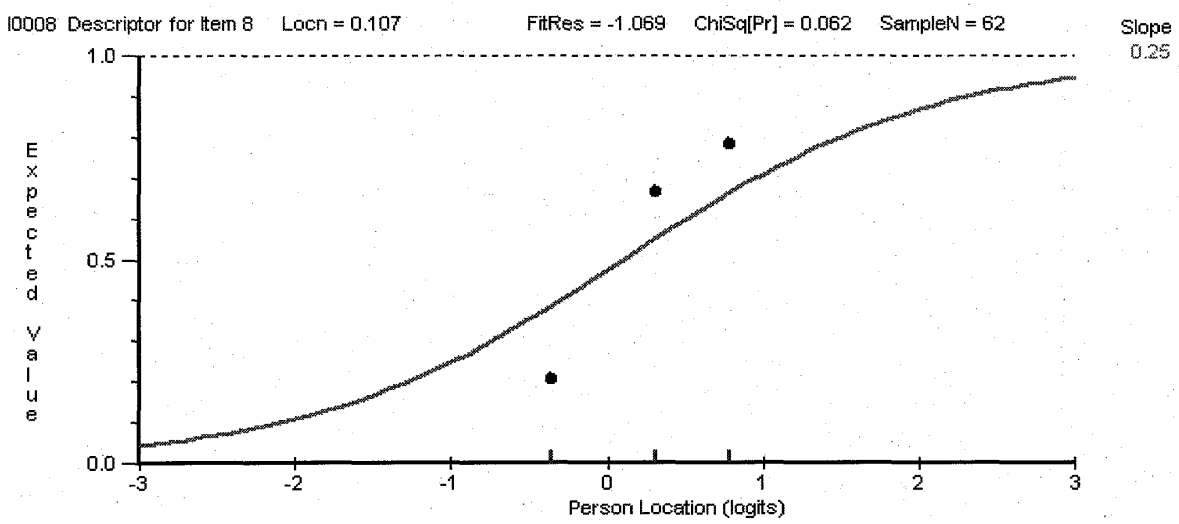
Type Three(item 8, item 10, item 16, item 18, item 19, item 24, item 26). The obtained curve of the three total score groups is so steep in distinguishing the three total score groups of persons.

4.4 Discussion of Differential Item Functioning

Let us now examine the Differential Item Functioning (DIF), that is, whether each item operates in the same way for each of the three groups of students, or majors (Note that these groupings are by major of Law, Business and Economics – not the three total score groups considered above in the section on Fit). We will check the ICC curves for each item from the viewpoint of the persons' majors by focusing on the Differential Item Functioning idea. Let us take a look at the following graphs.

Because of the small number in the sample, these curves tend to be all over the place! We would not show them all. Let us pick out five types with some

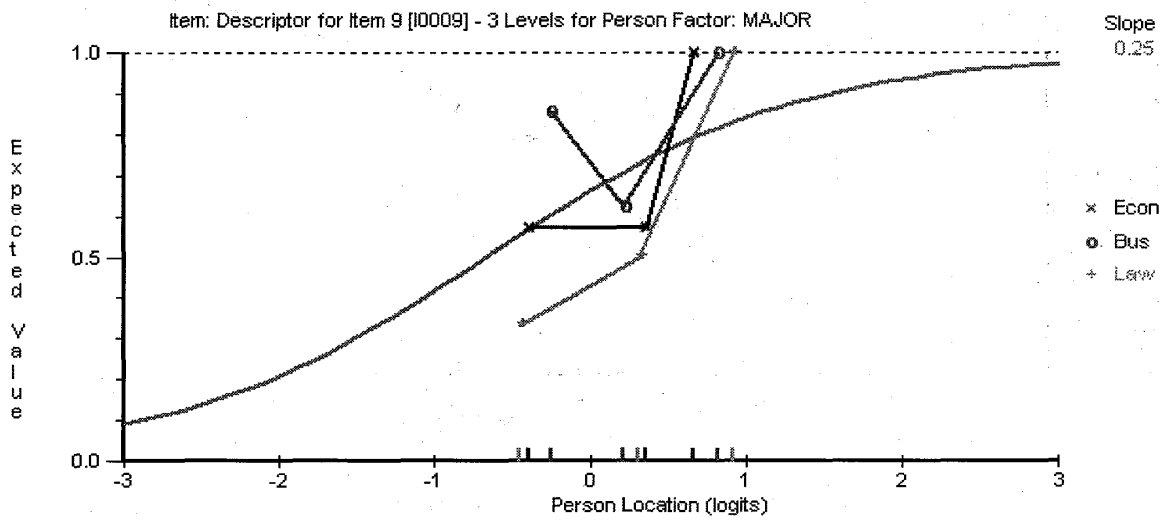
Figure 4.
ICC Graph of Type Three (e.g. item 8)



examples that illustrate DIF best.

Type One (items in which all three majors agree in their response patterns) has little or no DIF. Perhaps item 9 is the best of this type.

Figure 5.
DIF Graph of Type One (e.g. item 9)



Type two (items in which Law students=green line are performing idiosyncratically in their responses), We can use items 29 and 20 where Law is advantaged relative to the other two groups. We should examine the content of these items. Could we expect Law students to know this content better than students from other majors?

Type three (item in which Business persons = red line perform rather differently from the other two majors). We can show items 8 and 14 where Business is advantaged relative to other students. We should examine the content of these items.

Figure 6.
DIF Graph of Type Two (e.g. item 29)

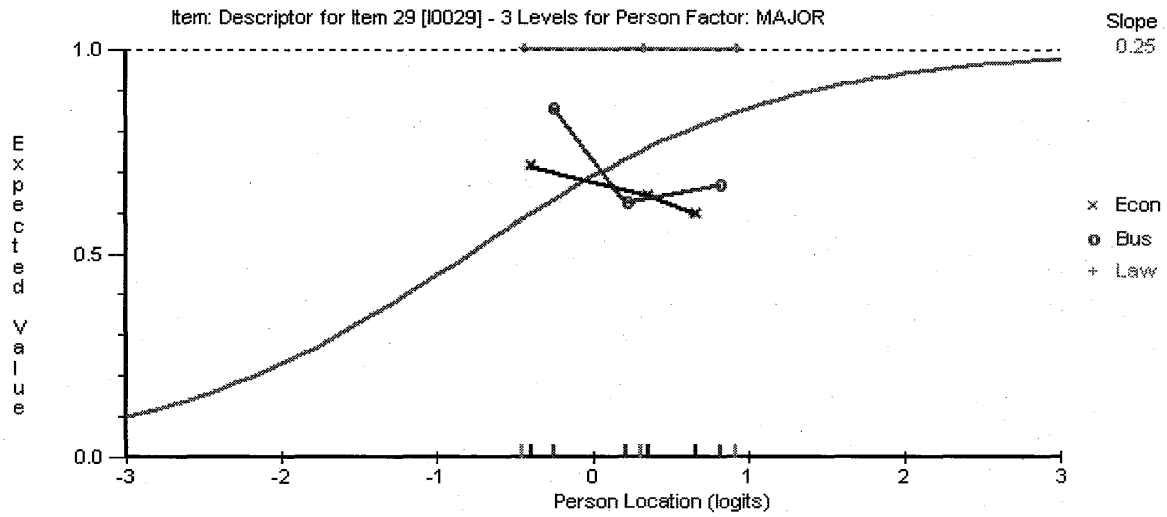
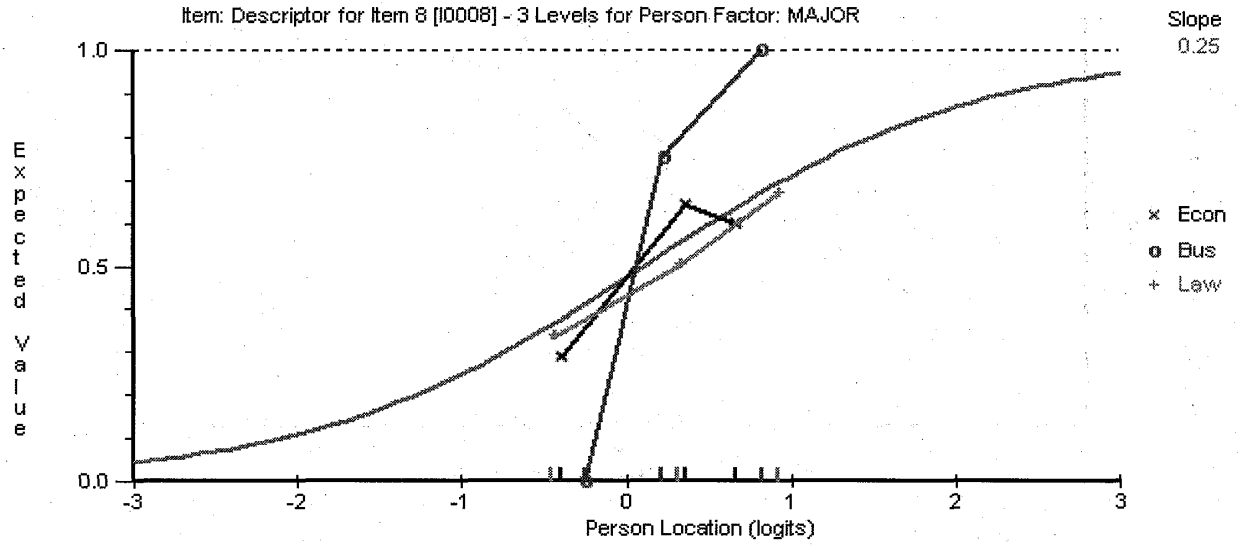


Figure 7.
DIF Graph of Type Three (e.g. item 8)



Type four (items in which Economics persons =blue line perform differently from the other two majors). We can show item 16 where Economics is advantaged relative to others. We should check the content of the item to see if there is any reason why this group would know this item better than the other two groups.

Also, items 22 shows that Business and economics are both advantaged compared with law. We need to examine the reason, based on the type of questions.

Figure 8.
DIF Graph of Type Four (e.g. item 16)

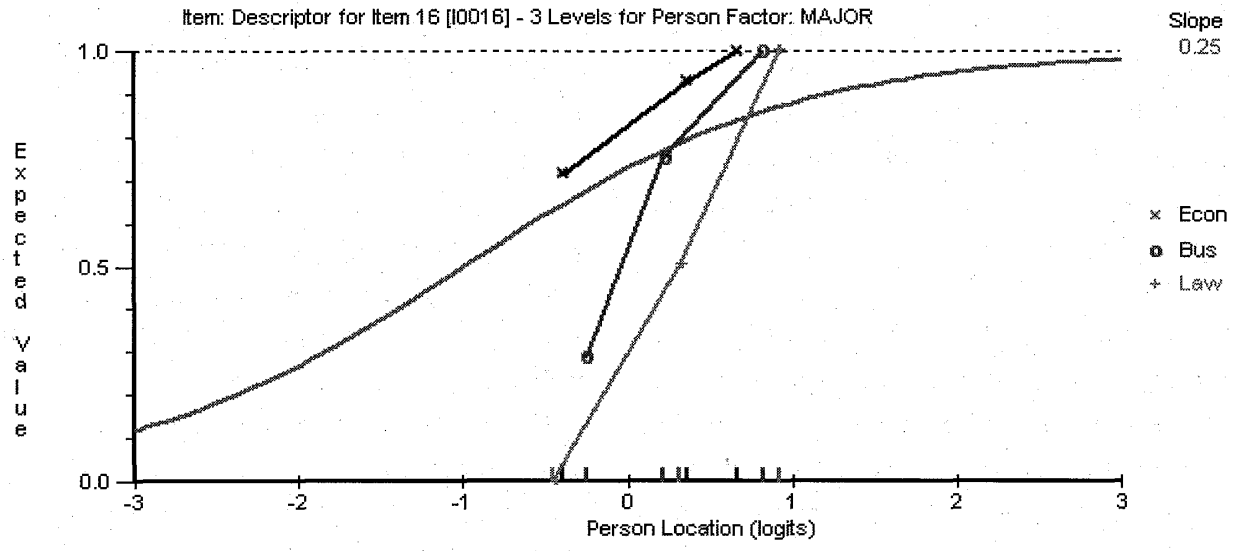


Figure 9.
DIF Graph of item 22

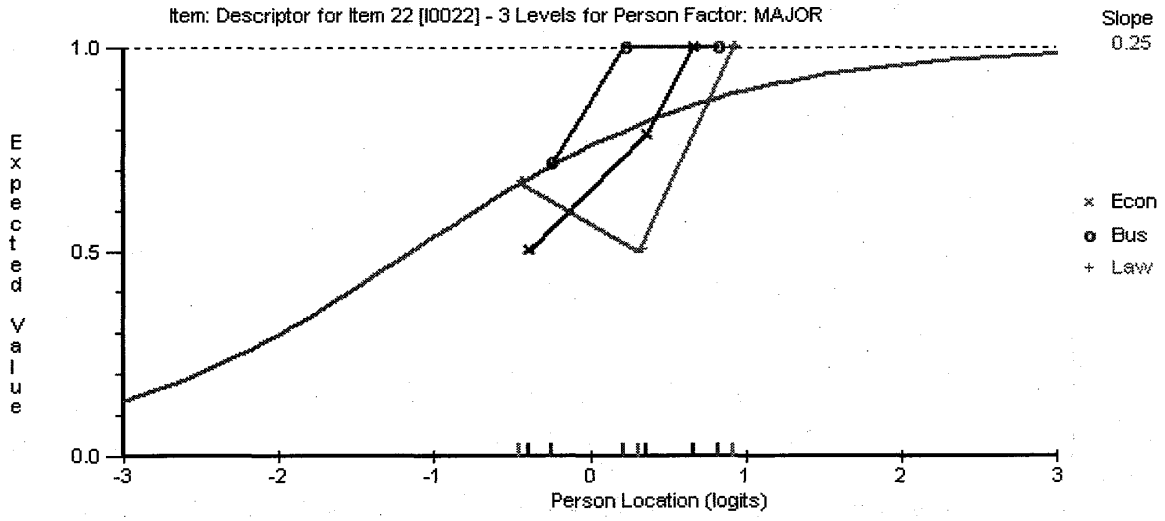
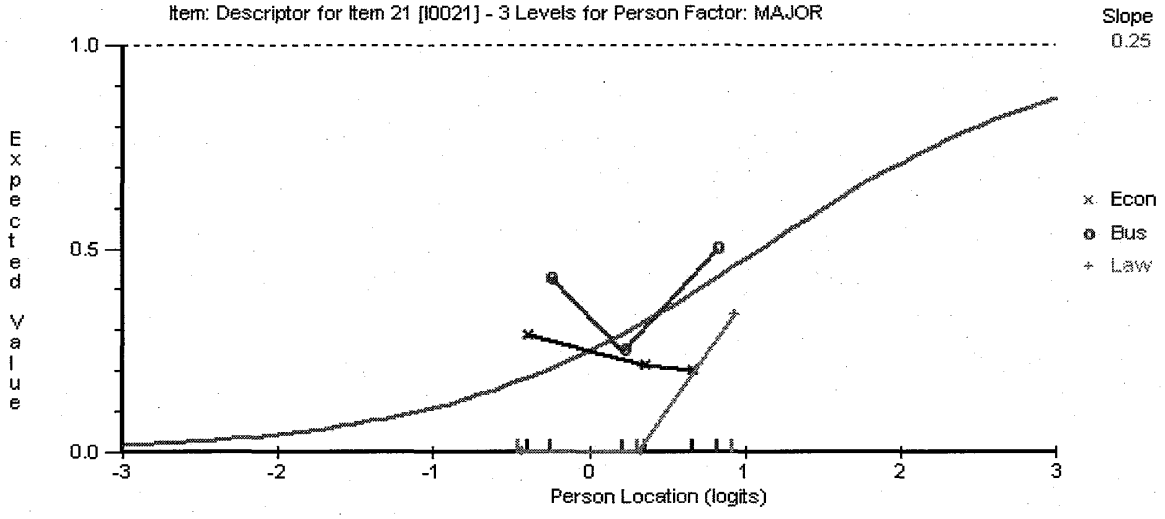


Figure 10.
DIF Graph of Type Five (e.g. item 21)



Type five (items in which all the three groups perform in individual ways) . We can show only item 21 here, because all items look a little inconsistent due to small numbers, so we need to pick one where the patterns for all three student groups are quite distinctive.

In summary, this above information from the DIF analysis shows for the three majors that Law students are often idiosyncratic so that the information from them is unpredictable However, the inconsistent patterns are likely to be due mainly to the fact that the sample for them is so small (8 students). More data is needed before decisions should be made based on their responses to this test. One reason for this can be caused by the too small number of the law students (only 8).

We can even explore the reason why all the Law students got item 29 correct, if the test-taker number is appropriate. In this way we can go into the details of the bias check through this DIF analysis.

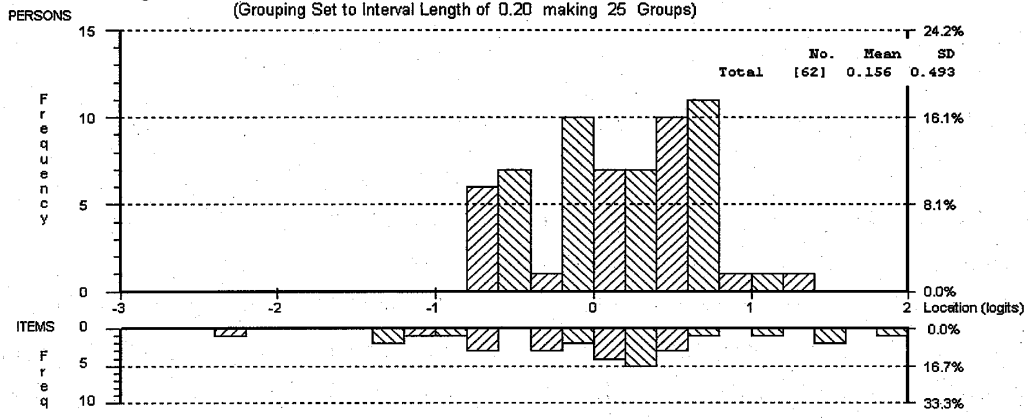
We can also point out that using Rasch model, we can check on validity, reliability but also how items function for different groups. Ideally they should all function in the same way, i.e. all people

with the same total score should have the same probability of getting any particular item correct (or incorrect). The point about DIF is that, for an item showing DIF, people from different groups (in this case, different majors), even when they have the same total score, have different probabilities of getting that particular item correct. i.e. the item is shown to be biased in favour or against them (depending which group you consider).

It would be helpful if we drew a line parallel to the y-axis on one of the pictures to show that people with the same total score (ie the same location on the scale overall) have different probabilities of getting an item correct when the three obtained curves for the three majors show DIF.

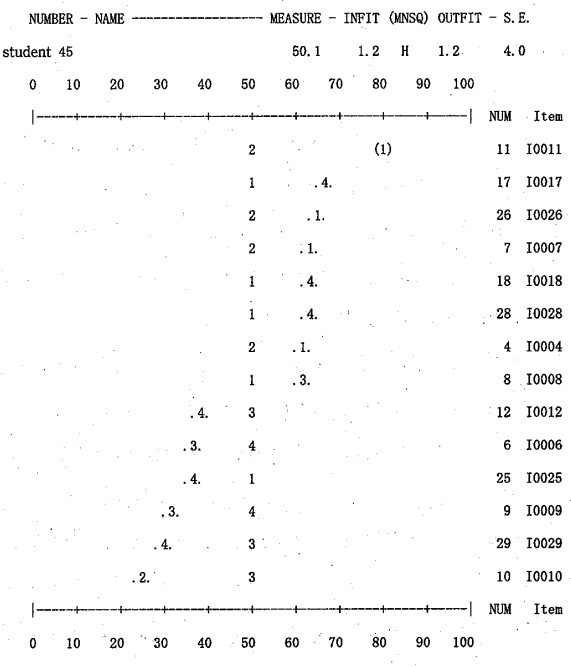
Targetting is used to check the distribution of people compared with the distribution of items. In this case, the scale could include more difficult items that would measure very able students a bit better.

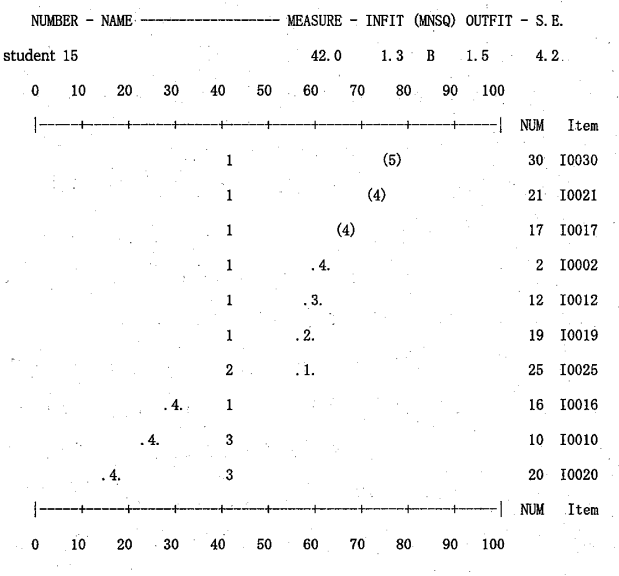
Figure 11. Person-Item Location Distribution

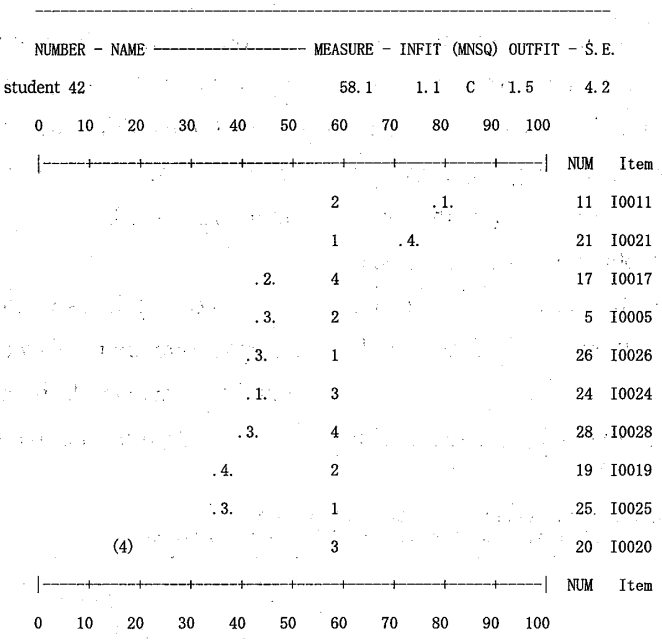


5 Analysis of Unexpected Response Patterns

Figure 12
Three examples of students' unexpected response patterns







Through the investigation of unexpected response patterns in terms of the respondents' observed scores, we are able to find a significant interaction between the students and the items, and eventually this result can be used for students' feedback and the improvement of the test.

In this study, students' abilities are shown in a 0-100 point scale which has been converted from -3 to 3 logit scores. Roughly speaking, in this study those students whose abilities (measures) are around 50 are operationally called intermediate, those students whose abilities (measures) are above 60 are operationally called good, and those students whose abilities are below 40 are operationally called poor.

Let us take a look at 3 examples in Figure 1. Student 45, whose measure (ability) was 50.1 (intermediate level) got a very difficult item (item 11) correct unexpectedly. The difficulty level of this item is far from the students' ability level. One possible reason for this could be lucky guessing.

Anther example is Student 15. This student' measure (ability) was 42 which is closer to the poor level, but got three very difficult items (item 30, item 21, item 17) correct unexpectedly against his/her true ability. One possible reason may be his/her lucky guessing.

Student 42, whose measure was 58.1 which is rather closer to the good level, got a very easy item (item 20) wrong unexpectedly. This could be caused by his or her careless mistake.

In this way, we can examine the unexpected answering pattern of students, and further explain the reason by interviews or giving questionnaires.

6 Conclusion

In conclusion, this paper provides a brief procedure of how thirty items that function differently for three groups of students (Business Majors, Economics Majors, and Law Majors) can be identified through the Rasch Model.

Traditionally, differences in performance between groups of students have been examined with the intention of identifying whether or not one group of students has been disadvantaged relative to another on traits of the test. However, through this DIF analysis it can be pointed out that groups of students perform differently on an item. The emphasis can be shifted from the tests to the items.

Through the investigation of unexpected response patterns of the respondents' observed scores, we were able to find significantly different interactions among the students on some items, and eventually this result can be used for the improvement of the test by asking students about their performance when giving feedback to students.

Acknowledgements:

This research was supported in part by Dr. Irene Styles and Dr. David Andrich of Murdoch University, Perth, Western Australia.

Bibliography

Andrich, D., Sherican, B., & Luo, G. (2004). *RUMM 2020: a Windows program for the rasch unidimensional measurement model.* Perth, Western Australia: RUMM Laboratory.

Andrich, D., & Styles, I. (2004). *Report on the psychometric analysis of the early development instrument (EDI) using the rasch model.* Centre for Learning , Change and Development. School of Education, Murdoch University.

University of Cambridge Local Examinations Syndicate. (1998). *Multilingual glossary of language testing terms: Studies in language testing 6.* Prepared by ALTE members. UK: Cambridge University Press.