Reprint of:

Shimizu, Y., & Zumbo, B. D. (2005). A Logistic Regression for Differential Item Functioning Primer. *Japan Language Testing Association Journal, 7,* 110-124.

研究紀要

JLTA Journal No. 7

第 7 号

2005

日本言語テスト学会
The Japan Language Testing Association

A Logistic Regression for Differential Item Functioning Primer

Yuko Shimizu
Ritsumeikan University

Bruno D. Zumbo
University of British Columbia

0. Introduction

The purpose of this article is to describe a statistical methodology, logistic regression (hereafter referred to as LogR), for differential item functioning (hereafter referred to as DIF) for language testing. We will illustrate LogR DIF with an English test used for a placement purpose with newly entered students in a private university in Japan. With an eye toward our purpose we will first provide a basic overview, including the definition, purpose, and methods of DIF. Second, to contextualize LogR DIF methods for language testers, we review some studies using DIF analysis in the field of language testing. We close with a step-by-step guide and demonstration of using LogR DIF statistical methods using a sample data of the aforementioned English test.

At this point, two are noteworthy. First, we focus on LogR as a statistical method for DIF analyses because as noted by Swaminathan (1994) LogR can be considered the most general form of the contingency table and generalized linear modeling approaches to DIF detection (Camilli & Shepard, 1994; Zumbo & Hubley, 2003). Second, it is these contingency table and generalized linear modeling approaches (and particularly the Mantel-Haenszel test) that are among the most widely used DIF statistical methods therefore LogR methods are a good foundation for building ones knowledge of DIF methods.

Therefore, given the lack of a literature on DIF among language testers in Japan, our goal is to illustrate how the statistical methodology of LogR DIF analyses can be a useful tool in test development and in establishing the validity of the inferences we make from our test scores. Readers interested in a more general overview of DIF methods should see Camilli and Shepard (1994), Clauser and Mazor (1998), and Zumbo and Hubley (2003).

1. Background and the Current Uses of DIF

Test fairness and test bias have been important topics in the field of testing and measurement in North America since 1960s. It is necessary to use a correct measurement tool that does not have bias in order to make decisions in various aspects in education including, but not limited to, screening and selection. Methods for detecting DIF and item bias are commonly used in the process of developing new measures, adapting existing measures, or validating test score inferences. DIF methods allow one to judge whether items (and ultimately the test they constitute) are functioning in the same manner in various groups of examinees. In broad terms, this is a matter of measurement invariance; that is, is the test performing in the same manner for each group of examinees?

In current practice there appear to be at least five distinct (but overlapping) purposes for DIF analyses.

1. *To insure fairness and equity in testing.* The groups are defined ahead of time here and often because of policy and legislation (e.g., visible minorities, gender, or language groups). As Zumbo and Hubley (2003) note, it is this purpose that was the rationale for the development of the earliest DIF methods. This use is most common in large-scale testing contexts wherein someone is using the test scores for decision-making, such as screening students for entry into jobs or entry into a college/university with a language test. Concerns about item bias emerged within the context of test bias and high-stakes decision-making involving achievement, aptitude, certification, and licensure tests in which matters of fairness and equity were paramount. Historically, concerns about test bias have centered around differential performance by groups based on gender or race.

2. *As evidence during litigation.* DIF analyses can be used as evidence in the case when failed candidates file a claim of discrimination because of unfavorable test results. In short, the DIF results, in particular, that show no DIF, would reduce the risk of litigation based on discrimination.

3. *As a statistical method to investigate if items are changing in terms of their difficulty and discrimination over time.* In large-scale testing programs that use the same items over-and-over-again over time the question of whether the items' difficulty and discrimination change over time is a question of concern. This is often referred to as "item drift". DIF analyses can be used as a statistical method to investigate whether items have maintained their psychometric properties over repeated use.

4. *Investigating DIF so that one can make group comparisons and rule-out measurement artifact as an explanation for the group difference.* This purpose is, in essence, about dealing with a possible "threat to internal validity" of group comparisons. The groups here are identified ahead of time and are often driven by research questions that the investigator has (e.g., gender differences in language test performance).

5. *Understanding the process of test responding.* There has been recent interest in investigating the cognitive processes of item responding and test performance, and investigating whether these processes are the same for different groups of individuals. In this context the groups need not be identified ahead of time and instead latent class or other such methods are used to "identify" or "create" groups and then these new

"groups" are studied to see if one can learn about the process of responding to the items. Of course, one can also investigate the process of responding for intact groups such as gender.

Regarding the various purposes of DIF analyses it is useful to keep two points in mind. First, we are now in what Zumbo (2004) has called the "Second Generation of DIF Research". Building on observations in Zumbo and Hubley (2003) this Second Generation DIF is no longer just rooted in matters of avoiding litigation or just flagging potential biased items. Today, as we can see from the list above, in addition to matters of bias, DIF technology is used to help answer a variety of basic research and applied measurement questions wherein one wants to compare item performance between or among groups when taking into account the ability distribution. At this point, applications of DIF have more in common with the research methodology aligned with analysis of covariance (ANCOVA) or attribute-by-treatment-interaction (ATI) than test bias per se.

This broader application has been the impetus for a variety of current and future directions in DIF development, such as test translation and cross-cultural adaptation -- see Zumbo (2003) for a language testing description. Many novel applications of DIF occur because previous studies of group differences compared differences in mean performance without taking into account the underlying ability continuum. An example of such an application in language testing would be a study of the effect of background variables such as discipline of study, culture, and hobbies on item performance.

Second, although it is rather fashionable these days to criticize DIF analyses for not providing the reason for differential test performance. It is clear from the above description of the purposes of DIF that that criticism is somewhat misplaced because not all DIF studies are aimed at finding the "reason" for DIF. One could, for example, only be interested in flagging DIF items in an operational language test, and hence the "reason" for DIF is secondary to guaranteeing the adequacy of the inferences made from the test scores, and hence reducing test bias against sub-groups of test-takers.

Previous studies in test bias have focused on differential performance by groups such as gender. For example, Lumsden and Scott (1987) conduced t-test and multiple regression to conclude performance differences between male and female students on essay tests and multiple-choice tests in the context of economic education. A more sophisticated approach to test bias in economics education can be seen, for example, by Walstad and Robson (1997) and Barrett (2001) using a DIF technique. In the field of language testing, a summary by Kunann (2000: 6) indicates that DIF analysis in terms of test fairness appeared in the beginning of 1980s, a main focus being gender as well as differences in test-takers' first languages.

2. Differential Item Functioning

The DIF analysis is a statistical procedure to determine if test items are appropriate for measuring the knowledge of various sub-groups of test takers. The test-takers background can define these sub-groups, for example. An assumption behind DIF is that test takers who have similar abilities, as indicated by the test score, will perform in similar ways on individual test items without being affected by the test takers' background; race, gender or ethnicity, for example. If particular items function differently for specific groups of test takers, they may reflect a bias that is not related to the domain being tested. This will consequently become a source of error in measurement. It is important to note, however, that DIF is not synonymous with bias. DIF is at statistical technique to detect differential performance on test items. That differential performance may be an artifact of the measurement process tapping some secondary and confounding latent variable, at which point it is suggestive of item bias. However, the observed differential item performance may be attributable to a process that is of interest to the test user (hence reflecting "true" differences rather than an artifact) resulting in what is called item impact, which will give us important implications for a learning-teaching process. DIF is therefore a necessary but not sufficient condition for item bias.

2.1 The Variety of DIF methods

Zumbo and Hubley (2003) describe three frameworks for DIF methodologies: (1) modeling item responses via contingency tables and/or regression models, (2) item response theory, and (3) multidimensional models. Although these frameworks may be seen as inter-related, they are freestanding. Each framework provides useful organizing principles for describing DIF and developing methods for detecting DIF in items. We will focus our discussion on the first framework because it is a flexible, does not require large numbers of items, and is easily adapted and implemented with widely available computer software; all of these being strengths in the language testing context.

2.2 Generalized Linear Regression Models for DIF Detection

A statistical implication of the definition of DIF (i.e., persons from one group answering an item correctly more often than equally knowledgeable persons from another group) is that one needs to match the groups on the ability of interest prior to examining whether there is a group effect. That is, the definition of DIF implies that after conditioning on (i.e., statistically controlling for) the differences in item responses that are due to the ability being measured, the groups still differ. Thus, within this framework, one is interested in stating a probability model that allows one to study the main effects of group differences (termed 'uniform DIF') and the interaction of group by ability (termed 'non-uniform DIF') after statistically matching on the test score. The item response is the dependent variable (i.e., the response variable) in the regression and the conditioning and grouping variables are the independent (i.e., explanatory or predictor) variables. This probability model can be stated in the most general sense as a generalized linear regression model, thus allowing for binary, rating scale, or continuous item responses.

The common regression approaches, and their advantages, are described for binary item data in Swaminathan and Rogers (1990) and for Likert or rating scale items in Zumbo (1999). This approach entails fitting a generalized linear regression model (in the most common case, a LogR analysis as the scores are binary) for each item wherein one tests the statistical effect of the grouping variable(s) and the interaction of the grouping variable and the total score after conditioning on the total score.

In short, one fits two logistic models to the data, and compares the difference in the -2 log likelihoods of the first and second models to a χ^2 distribution with 2 degrees of freedom. The first model included terms for the ability of each respondent:

$$\text{Logit p(item response is correct)} = \beta_0 + \beta_1{*}\text{ability} \qquad (1)$$

and the second models adds a grouping variable (denoted group) and a term for the group-by-ability interaction:

$$\text{Logit p(item response is correct)} = \beta_0 + \beta_1{*}\text{ability} + \beta_2{*}\text{group} + \beta_3{*}(\text{group}{*}\text{ability}) \qquad (2)$$

This second model is compared to the model that included only the ability term (model 1). The difference in the -2 log likelihoods of these two models was compared to a χ^2 distribution with 2 degrees of freedom. If the χ^2 (2 df) statistic was statistically significant at α=0.05 level, that item was flagged as exhibiting DIF. It should be noted that the regression coefficient for the grouping variable reflects uniform DIF and that for the interaction reflects non-uniform DIF. Therefore, using this 2-df strategy uniform and non-uniform DIF is thus identified in a single step. An alternative strategy commonly used in practice is to report two 1-df tests for each of the uniform and non-uniform DIF; single degree of freedom Wald tests.

Several important points of flexibility arise. First, one can potentially match on more than one variable. Second, likewise, one can easily have more than two groups. Third, of course, if one has continuous item responses one can simply apply regular ordinary least-squares regression if it is appropriate. Fourth, one may be concerned with the reliability (i.e., measurement error) of the matching variable and hence one may either:

(a) estimate a latent variable score and match on it,

(b) purify the matching variable by two-step strategy wherein one runs a first pass at a DIF analysis and then removes any items detected as DIF in this first pass from the total score; the second run of DIF analyses on the items then conditions on this newly created "purified" total score, or

(c) one can compute a "rest" score (much like item-total correlations) for each by creating a total score without the item under study and then matching on this "rest" score for the DIF analyses.

Finally, a natural extension of this methodology is for mixed effects (i.e., hierarchical linear modeling HLM or multi-level modeling) regression models that can incorporate complex data structures arising from educational or survey data. These extensions are rather straightforward to apply but they do add complexity to the interpretation of the parameters.

3 DIF Approach in Language Testing

Research on bias in language testing has been directed towards test takers' differential performance on various types of tests that would be attributed to their gender and language background. While analysis of variance gives an answer for such a phenomena, a use of DIF techniques that match on the test score performance has been of interest since mid-1980s. Application of the DIF technique to language tests was pioneered by Alderman and Holland's (1981) examination of the Test of English as a Foreign Language (TOEFL), followed by Chen and Henning (1985)'s study of different language groups. Chen and Henning used Rasch Model difficulty estimates analogous to the delta-plot technique proposed by Angoff (1973) to examine DIF on a well-established placement test battery used at a university in the USA with two language groups of Chinese (n=77) and Spanish (n=34) among more than 30 language groups. They concluded that the vocabulary test was the source of Chinese/Spanish bias. However, the sample size was not large enough for the difficulty parameter for reliable calibration.

Ryan and Backman (1992) also examined the DIF among different language groups. They analyzed the results of the First Certificate of English (FCE) and the TOEFL to detect DIF across Indo-European (n=792) and Non-Indo-European (n=632) language groups, using the Mantel-Haenszel procedure for their analysis. The results showed that 32 items out of 146 items were differentially easier for Indo-European groups and 33 easier for Non-Indo-European -- the distribution being varied in three sections of the TOEFL. In the FCE, the similar results were observed. That is, 25 items out of 40 showed DIF among the two language groups. In this study, gender difference was not observed, that is, only with one structure item and three vocabulary items were found to be easier for male test takers.

More recently, a DIF analysis between Asian and European test-takers was reported by Kim (2001) focusing on (a) short tests, and (b) a speaking test in which the responses were scored polytomously. The methods used were the likelihood ratio tests and Zumbo's ordinal LogR procedure (Zumbo, 1999). A total of 1038 subjects from six different countries took a speaking test which consisted of six tasks: reading aloud, sentence completion, picture sequence, single picture, free response questions and telling about a schedule in a tape-mediated format. Four-point scale (0-3) was used to assess grammar, pronunciation and fluency. The results showed that the grammar scale and pronunciation scale of the given test functioned differently across the Asian and the European language groups. Also, group membership had some effect on low to middle level test-takers' performance, which implied that DIF was more sensitive to those levels than the high ability level. The sources of DIF were then examined in terms of the test characteristics and the test-takers' characteristics and indicated possible influence of the types and the number of scoring scales onto the test validity.

In a homogeneous situation of foreign language education, wherein the learners share the common first language, variables other than the language backgrounds and cultural differences among test-takers are a great concern. The extensive study on second language (L2) vocabulary was conducted with Finish test-takers by Takala and Kaftandjieva (2000) to study gender impact. They analyzed gender DIF in a L2 vocabulary test, English in this case, and potential gender impact on the performance observed by different item composites. The vocabulary test, which was a part of the test battery of the Finnish Foreign Language Certificate Examination contained 40 multiple-choice items and was taken by 182 males and 293 females. The data was analyzed with the One Parameter Logistic Model (OPLM). Eleven items were found to differ in their functioning in terms of difficulty in favor of either males or females. Although the observed differences in the results remained even after excluding those DIF items, they pointed out the need of more empirical verification and suggested exclusion of DIF items from item bank building point of view.

In the language environment in higher education, students' specialization arouses great interest. Henning (1990) focused on DIF attributed to academic specialization of the test takers. Four specialized groups of Physical Science (n=56), Humanities (n=38), Fine Arts (n=36), and Business (n=57) were formed with entering graduate and undergraduate students and their performances on an English placement test, which consisted of five subtests, were analyzed using the Angoff Group Item Difficulty Scatterplot Method, the Regression Residual Method and the Mantel-Haenszel Method. Three subtests of listening, reading and error detection were made of passages with topics relevant to the respective specialization categories of the test takers, while two subtests of grammar and vocabulary were not labeled under any specialization categories. The results indicated that the passage-bound items were not identified to be biased in favor of any specialization except three items. That is, no systematic specialization bias was observed for or against persons who worked in particular academic fields based on passage content selection. Henning suggested that the inability to detect specialization bias might result form the possibility that the test and the background of the test takers were bit specialized enough.

Pae (2004) also highlighted academic specialization of test-takers. Using the Item Response Theory Likelihood Ratio approach, s/he identified DIF on the English subtest of the Korean National Entrance Exam for Colleges and Universities for test takers with different academic backgrounds. A randomly selected sample in this study consisted of 14000 college-bound high school seniors: 7000 Humanities and 7000 Sciences. The English test was comprised of 55 multiple-choice items: 17 items to measure Listening Comprehension (LC) and 38 items for Reading Comprehension (RC). The results identified 18 DIF items across the subscales. Five items in LC subscale were flagged for DIF, three of which flagged for non-uniform DIF were differently more difficult for the Sciences. In RC items, among the 13 DIF items, nine items were non-uniform DIF, were more discriminating for the Humanities. By academic groups, seven items (3 LC items and 4 RC items) were easier for the Humanities and nine items (2 LC items and 7 RC items) were differently easier for the Sciences. Concerning the relationships between those DIF directions and items contents, DIF might be assessed by content analysis; however, item content alone was not a reliable predictor.

Finally, Zumbo (2003) in the context of translation of language tests, showed that scale-level comparisons among groups via factor analyses or invariance tests in structural equation modeling alone do not allow one to detect DIF or item bias. That is, conducting scale-level analyses are not enough to ensure that DIF has been detected; one needs DIF analyses.

Clearly, there has been a longstanding interest in DIF methods in language testing, primarily for the purpose of test fairness. There are few consistent and programmatic findings but perhaps it is too early in the program of research to expect such. So that we can foster more programmatic DIF research in language of the three purposes described in section one, let us turn to a demonstration of how one can use commonly available statistical software to conduct DIF analyses.

4. Demonstration with Language Testing Data

In the current study, we analyzed a sample data set of an English test administered for a placement purpose for first year students at a university level. The test was revised after conducting a pilot test and classical item analysis in advance of the actual administration. By showing the DIF technique with this data set, we underscore the need for on-going validation in the light of test bias analysis as a part of the test development process.

University admission in Japan is based largely on the scores achieved by students in entrance examinations, which include English tests in many cases. However, for many universities, private ones in particular, the newly enrolled students in the same academic year, in the same institute, do not necessarily take the same examinations or parallel tests. Therefore, there is no way to grasp the students' English proficiency under the common measurement scale in the same university or program unless some tests are conducted after their entrance. Some universities adopt commercially available English tests if necessary. Or some, but a few, develop an English test that is more appropriate for their target students and condition. In this study, we use a data set of an English test developed for a placement purpose at a private university to illustrate the statistical technique for DIF analysis.

Note that the most common grouping variables in language testing DIF analyses are (a) country of origin or cultural background, and (b) gender. As Zumbo and Hubley (2003) note, many novel applications of DIF occur because previous studies of group differences compared differences in mean performance without taking into account the underlying ability continuum. An example of such an application in language testing would be a study of the effect of background variables such as discipline of study on item performance. With this in mind we focused on faculty of study, Economics (EC) and Science and Technology (ST) as our grouping variable because group differences in test performance have been found in the past and, as described above in section 1, one needs to rule out a measurement artifact in comparing the groups. In addition, matters of test fairness and equity also motivated us.

4.1 Instrumentation

The data used in this study are form the results of a placement test conduced with first year EC and ST students at a private university in Japan. In this university, they developed the English Placement Test (hereafter PLT) that has been administered to first-year students every April (i.e., the beginning of the school year). The main purpose of the PLT is to make teaching-learning of the required English program more effective by placing the learners into their suitable English proficiency levels. Note that the PLT was developed not to label or classify students using particular specialization rubrics, but to ascertain the students' general English ability that is based on what they achieved at high school.

The test battery consists of three subtests of listening, grammar and reading requiring about 60 minutes to administer. Each subtest has 30 multiple-choice items. The detailed test structure is show in Table 1. We will be focusing on the grammar subtest for our demonstration.

Table 1 Structure of the PLT

Subtest	Main tasks	Number of items		Time allotment (minutes)	Score
I. Listening	Question-response	15	30	20	60
	Dialogue	10			
	Short lecture	5			
II. Grammar	Fill-in the blanks	20	30	15	60
	Error identification	10			
III. Reading	Cloze passage	15	30	25	60
	5 short readings	15			
Total		90		60	180

Although the test items were carefully written and screened for content not to favor students with different academic backgrounds, ST and EC students in this case, the former performed significantly better than the latter in some subtests, although not consistently from year to year. This might simply result from chance differences among entering students in a given year. An alternative explanation is that there may be test bias or impact. The test was initially developed using classical item analyses and hence the question of DIF did not arise. Therefore further validation research using DIF analysis is indispensable as a part of test development and understanding the construct.

4.2 Participants

The current study used data from the PLT administered in April 2003. A total of 2372 first year students took the PLT during the Freshman Orientation Program in April. Of these, 912 students were in Economics and 1460 in Science & Technology. Overall, nearly 26% of the EC students and 10% of the ST students were female.

4.3 Results

Because it is widely used in practice, the single-degree-of-freedom binary LogR DIF tests, for uniform and non-uniform DIF separately, were conducted for each of the 30 grammar items of the PLT. Each of these 1-df tests were conducted at an α=0.025 level (i.e., α=0.05/2 for the two DIF hypothesis tests; this maintains the item level α at less than or equal to 0.05).

Recall that the purpose of this paper is to illustrate the use of a statistical method, LogR, for DIF analyses. If we were reporting, and hence focusing on the DIF results, per se, we would go through the following steps:

1. Conduct a LogR DIF analysis for each item.
2. Report a table listing how many, and which, items did and did not show DIF.

3. Report a table listing the item-by-item DIF results including the odds ratios and other effect size measures.

4. Conduct a content analysis that can serve to help explain (i.e., provide possible reasons) for the DIF that was, and was not, found. In the case of the current data example we would also explore how the specialized knowledge of the two majors may allows us to understand when and why DIF may occur.

5. To aid in the investigation of possible sources, explanations, and reasons for DIF, we would recommend creating a new data set that has the item level information so that each item would be a row in the data matrix and the columns (i.e., variables) would be (a) whether the item was flagged as DIF, and (b) a series of columns that codify the results of the content analysis. This data sheet could then allow one to model and test hypotheses about potential DIF explanations. The important point is that the unit of analysis would now be the item.

The focus of this paper is step one. We focus on the statistical method because it is the one step in the above five that is most foreign to language testers. With this in mind, we will use items that tap "detection of similar meaning" (items 33 and 50 of the test) as an example of DIF. The 1-df Wald tests were both statistically significant, uniform DIF Wald(1 df)=5.9, p=.016 and non-uniform Wald(1 df)=6.5, p=.011. Figure 1 is the predicted logistic curves (based on the DIF regression models) for item 33. As an interpretative tool for the non-uniform DIF, one can see from Figure 1 that the item response curves cross at a total score of approximately 15 (out of 30 possible points). Clearly, at the lower end of the score distribution Science and Technology students do better than the Economics students but that this is reversed for scores greater than 15.

Yet another way of looking at this visual display in Figure 1 is that the item discriminates among test takers better for Economics students than Science and Technology students. It is important to note, however, that the item is not a particularly good one because it does not discriminate well among examinees (i.e., the slope of the curve in the mid-range is rather flat) and the lower asymptote of the curve is substantially greater than .50 for both groups. As a relative statement, however, the item performs relatively worse for Science and Technology students than Economics students. As an aside, it is interesting to note that the 2-df Chi-square test of DIF was also statistically significant, χ^2 (2 df)=6.5. p=.038.

In contrast to item 33, item 50 had no DIF. The 1-df Wald tests were both statistically non-significant, uniform DIF Wald(1 df)=0.50, p=.48 and non-uniform Wald(1 df)=1.6, p=.20. Figure 2 is the predicted logistic curves (based on the DIF regression models) for item 33. One can see in Figure 2 that the predicted item response curves are nearly coinciding which means that the item is functioning the same way for both groups, i.e., no DIF. In fact, item 50 is an ideal item for test functioning near the middle of the distribution because the 50:50 threshold appears to be at about 17, and the item response function tends toward zero at the lower end of the score distribution.

5 Closing Comments

What follows are some closing comments on DIF and LogR methodology.

First, a question that sometimes arises when one uses any statistical DIF methodology is the build-up in Type I error rate because of the multiple hypothesis testing. One must recall that a Type I error occurs when one rejects the null statistical hypothesis (in this case no DIF) when in fact one should not have done so -- in this case there as actually was no DIF but we declared an item as DIF. The error of doing so results in an item being scrutinized for sources of DIF in a post-statistical content analysis by subject-matter-experts and content specialists when it is actually not DIF. There are no simple answers as to how one should handle setting α for the DIF hypothesis tests. Instead, one needs to weigh the error of various alternatives and select a Type I error rate that fits the overall purpose of the test (taking into account the severity of an error in overall test score due to item bias).

Second, although the focus of analysis is each item, in the end one needs to reflect on what the DIF means for the test as a whole. In this case, our example may be a useful demonstration of how one might approach this. In repeating the steps in the LogR analysis we demonstrated above for each item in our example test only three of the 30 items showed DIF. Given that 90% of the items showed no DIF the analytical results in Rupp and Zumbo (2003) suggest that inferences from the overall test score would not be compromised because for so few items Rupp and Zumbo show that the overall test score is not greatly affected.

And in closing, DIF results can be used to create working hypotheses that fuel an on-going program of research for any test, be high-stakes or not. That is, again focusing on all of our items, rather than just the demonstration above, where DIF was found, the items tended to require memorization and local understanding such as detection of similar meaning, adverbs, and relative clauses. These findings would lead a testing organization to do follow-up research with a large item pool to test the conjecture or working-hypothesis about the item types. Of course, testing organizations need to feedback the results of DIF analyses to item writers so that eventually, over the course of test development and refinement, item writers and others in the testing organization can get a better sense of how items function for various groups of examinees.

Figure 1.
LogR DIF Plot for Item 33 (DIF item).

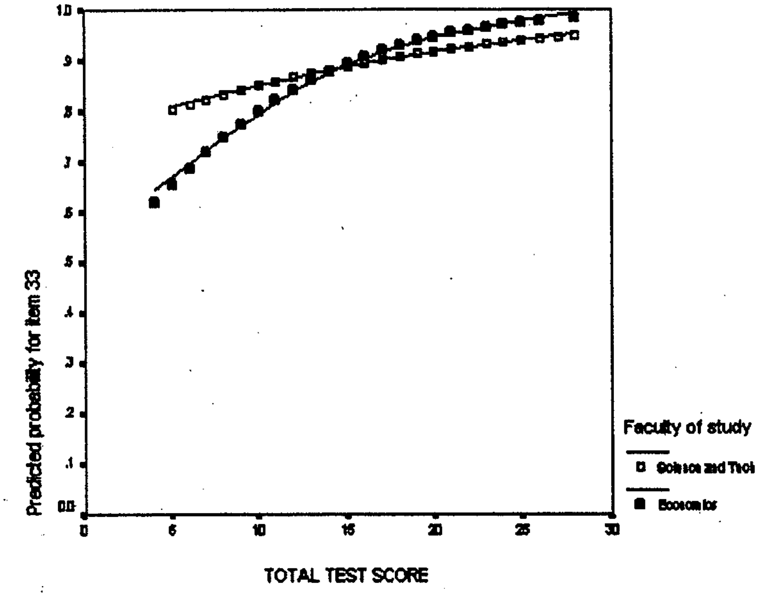
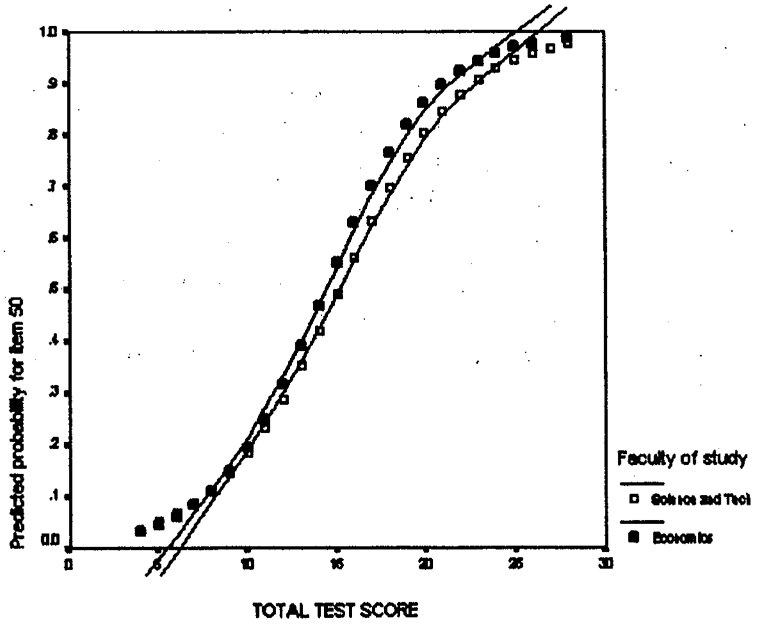


Figure 2.
LogR DIF Plot for Item 50 (No DIF).



Authors' Note: This project was conducted while Yuko Shimizu was a visiting scholar at the University of British Columbia in Professor Zumbo's lab. Information about the Edgeworth Lab at UBC can be found at the following URL http://www.educ.ubc.ca/faculty/zumbo/zumbo.html

Reference

Alderman, D.L. and Holland, P.W. (1981). *Item performance across native language groups on the Test of English as a Foreign Language*. Princeton, NJ: Educational Testing Service.

Angoff, W.H. and Ford, S.F. (1973). Item-rate interaction on a test of scholastic aptitude. *Journal of Educational Measurement, 10*, 95-106.

Barrett, S. (2001). Differential item functioning: A case study from first year economics. *International Education Journal, 2*, 123-132.

Chen, Z. and Henning, G. (1985). Linguistic and cultural bias in language proficiency tests. *Language Testing, 2*, 155-163.

Clauser, B. E., & Mazor, K. M. (1998). Using statistical procedures to identify differentially functioning test items. *Educational Measurement: Issues and practices, 17*, 31-44.

Henning, G. (1990). National Issues in Individual assessment: The consideration of specialization bias in university language screening tests. In de Jong, J.H.A.L. and Stevenson, D.K. (Eds), *Individualizing the assessment of language abilities*.(pp.38-50). Clevedon, Avon: Multilingual Matters Ltd.

Hubley. A. M., & Zumbo, B.D. (1996). A dialectic of validity: Where we have been and where we are going. *The Journal of General Psychology, 123*, 207-215.

Kim, M. (2001). Detecting dif across the different language groups in a speaking test. *Language Testing, 18*, 89-114.

Kunan, A.J.(ed) (2000). *Fairness and validation in language assessment*. Cambridge: Cambridge University Press.

Lumsden, K.G. and Scott, A. (1987). The economics student reexamined: male-female differences in comprehension. *Research in Economic Education, 18*, 365-375.

Ryan, K. and Bachman, L.F. (1992). Differential identity functioning on two tests of EFL proficiency. *Language Testing, 9*, 12-29.

Pae, T. DIF for examinees with different academic backgrounds. *Language Testing, 21*, 53-73.

Rupp, A. A., & Zumbo, B. D. (2003). Which Model is Best? Robustness Properties to Justify Model Choice among Unidimensional IRT Models under Item Parameter Drift. {Theme issue in honor of Ross Traub} *Alberta Journal of Educational*

Research, 49, 264-276.

Swaminathan, H. (1994). Differential item functioning: A discussion. In Dany Laveault, Bruno D. Zumbo, Marc E. Gessaroli, and Marvin W. Boss (Eds.), *Modern Theories of Measurement: Problems and Issues.* Ottawa, Canada: University of Ottawa.

Swaminathan, H. & Rogers, J. (1990). Detecting differential item functioning using LogR procedures. *Journal of Educational Measurement, 27(4),* 361-370.

Takala, S. and Kaftandjieva, F. (2000). Test fairness: a DIF analysis of an L2 vocabulary test. *Language Testing, 17, 3,* 323-340.

Waslstad, W.B. and Robson, D. (1997). Differential item functioning and male-female differences on multiple-choice tests in economics. *Journal of Economic Education, 28,* 155-171.

Zumbo, B.D. (1999). *A handbook on the theory and methods of different item functioning (DIF): LogR modeling as a unitary framework for binary and likert-type (ordinal) item scores.* Ottawa. ON: Directorate of Human Resources Research and Evaluation, Department of National Defense.

Zumbo, B. D. (2003). Does item-level DIF manifest itself in scale-level analyses?: Implications for translating language tests. *Language Testing, 20,* 136-147.

Zumbo, B. D. (2004). *Psychometric Methods for Enhancing Fairness and Equity: Differential Item Functioning (DIF) and Scale-level Invariance.* Invited Presentation at the Fourth International Conference, Equitable Assessment Practices: Building Guidelines for Best Practices, International Test Commission Conference. Williamsburg, Virginia, USA.

Zumbo, B.D., & Hubley, A. M. (2003). Item Bias. In Rocio Fernandez-Ballesterons (Ed.). *Encyclopedia of Psychological Assessment* (pp. 505-509). Thousand Oaks, CA: Sage Press.