



Moderation of Interviewer Judgment in Tests of Spoken English

Patrick Griffin

Nathan Zoanetti

Assessment Research Centre

The University of Melbourne

Paper presented at the XXXx TESL conference

Tampa, March CC , 2006/

This study examined the way interviewers differentially judge response quality when administering an interview test. Two methods were employed. The first was an approach called differential item functioning and the second examines judgement rating patterns. The first examined the rating behaviour of individuals and the second examined the behaviour of groups. Its purpose was to introduce a procedure of quality control into an interview procedure.

Interviewers were trained in the implementation of the International Test of English. The test consists of a series of 23 items. Each item was scored on a 3 point scale using strict rubrics. The original concept for the test was published in 1986 (Griffin et al., 1986) employed verbal algorithms. These have been used to redevelop the test and to work towards an internet delivered version.

Wherever judgement is involved there is a possibility of different judgement rules being applied and different scores being assigned to candidates of equal ability and performance on the same item by different judges even when extensive training is involved. When this occurs systematically over groups of individuals, the effect is called differential item functioning (diff). Usually differential item functioning (diff) refers to a difference in item performance between comparable groups of examinees (Dorans & Holland, 1992). Test items exhibit differential item functioning (DIF) if the item scores of equally able examinees from different groups (e.g. of different race, sex, or age) are systematically different (Kelderman and Macready, 1990, p307). It refers to a psychometric difference in how an item functions across different groups. It also refers to a difference in item performance between comparable groups of examinees, that is, groups that are matched with respect to the construct being measured by the test. The two groups are typically referred to as the Reference (R) group and the Focal (F) group. However these definitions apply to the case where only two groups are compared. With recent developments in software such as Conquest, (Wu and Adams, 1997) these labels are redundant and their definitions can be broadened. The reference group can be thought of as a 'benchmark' for item behaviour and the focus groups can be any groups compared to that benchmark. In the

case of item response modelling (IRM) and its examination of diff, the modelled item characteristic curve becomes the benchmark and the focus groups are the groups of interest that are compared to this benchmark.

Method

The assigned scores for the interview assessment were treated as 21 separate test items scored using a partial credit approach, with the highest score for each item reflecting the highest quality performance for that item and the lowest score reflecting the lowest quality performance. Scoring each item in this manner treats them as 21 independent polytomous items, in which each student, n , has a spoken language proficiency θ_n and each item has a set of difficulty parameters $\delta_{i1}, \delta_{i2}, \delta_{i3} \dots \delta_{ik}$ representing the difficulty of attaining each of the scores from 1 to k for item i . Each of these parameters governs the likelihood of a student with ability, θ_n , being given a score of k rather than $k-1$. The analysis models the relationship between student writing ability and the difficulty parameters of each of the 21 items. The Rasch model estimates student spoken language ability independent of which particular items were used for the estimation. The natural logarithm of the odds of achieving a specific score of k rather than $k-1$ is obtained from the simple relationship

$$\ln \frac{p_k}{p_{k-1}} = \ln \frac{n_k}{n_{k-1}}$$

where p_k and p_{k-1} are the proportions of students scoring k and $k-1$ respectively.

Most items were scaled using IRT (Item Response Theory) scaling methodology. With the One-Parameter (Rasch) model (Rasch 1960) for dichotomous items, the probability of assigning a score of 1 instead of 0 is modelled as

$$P_i(\theta) = \frac{\exp(\theta_n - \delta_i)}{1 + \exp(\theta_n - \delta_i)} \quad (1)$$

where $P_i(\theta)$ is the probability of person n to score 1 on item i . θ_n is the estimated latent spoken language proficiency trait of person n and δ_i the estimated location of item i on this dimension. For each item, item responses are modelled as a function of the latent trait θ_n .

In the case of items with more than two (k) scoring categories (as for example with a maximum x =score greater than 1) this model can be generalised to the *Partial Credit*

Model (Masters and Wright, 1997)¹. The Partial Credit Model developed by Masters (1982) is an extension of the Simple Logistic Model, and overcame the restriction to dichotomous scoring. The model was developed by estimating parameters for the difficulties associated with a series of performance levels within each item. Masters (1982) argued that the difficulty of the k^{th} level in an item governs the probability of responding in category k rather than in category $k - 1$. The probability of person n of completing the k^{th} level is specified by Masters (1982; 158) as:

$$P(X_{ni} = x) = \frac{\exp \sum_{k=0}^x (\theta_n - \delta_{ik})}{\sum_{h=0}^{m_i} \exp \sum_{k=0}^h (\theta_n - \delta_{ik})} \quad (2)$$

The model estimates the probability of a person n scoring x on the m_i performance level of item i as a function of the person ability on the variable being measured and the difficulties of the m_i levels in item i . The observation x is a *count* of the successfully *completed* item levels, while only the difficulties of these completed levels appear in the numerator of the model. The model provides estimates of person ability θ_n and item step level difficulty δ_{ik} and $P_x(\theta)$ denotes the probability of person n to score x on item i . θ_n denotes the person's position on the latent trait, the item parameter δ_{ik} gives the location of the item step, k , on the latent continuum and denotes an additional step parameter.

The IRT approach also enables an assessment of how closely the obtained data is predicted by the mathematical model. The predicted data is called the modelled data and the comparison of observed and modelled data allows an examination of fit.

¹ An alternative is the Rating Scale Model (RSM) which has the same step parameters for all items in a scale (see Andersen, 1997).

Item fit was assessed using the weighted mean-square statistic (infit), which is a residual based fit statistic. Weighted infit statistics were reviewed both for item and step parameters. The ACER Conquest software (Wu, Adams and Wilson, 1997) was used for the estimation of item parameters and the analysis of item fit.

Given that the 21 items (or sub-tasks) had variable maximum scores, the partial credit model (Wright & Masters, 1982) using the computer program ConQuest (Wu and Adams, 1998) was used to derive the estimates of item difficulty and student writing ability.

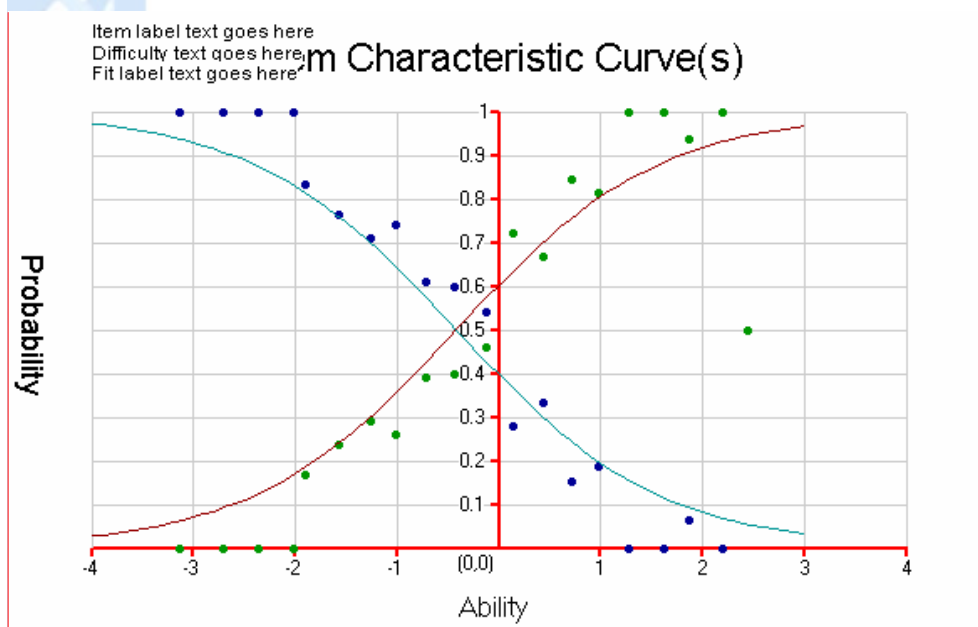
Under an IRM framework, a test item was showing diff if the Item Characteristic Curve (ICC) was not the same for the groups being assessed and if these were different in important ways to the benchmark curve; that is candidates who are equal in terms of the latent trait or ability do not have the same probability of being assigned the same score on the test item (Embretson & Reise, 2000). Diff is the effect when candidates who are equal in terms of the ability being measured by a test come from different subgroups and in general membership of the subgroup systematically affects the probability of being assigned a specified score (Camilli & Shepard, 1994). Membership of a group is therefore a determining factor in terms of scores on the test.

Typically, classical methods use an internal criterion such as total test score or "other items in the test" as the criterion for matching examinees to see if "comparable examinees from different groups" performed the same on individual test items.

It has been common to examine diff based on focus groups defined by gender, language background and place of residence or school type. In these cases it is not possible to change the way the grouping variable interacts with the item and it is common for the item to be discarded where diff illustrates important deviation from a benchmark or reference group. In this case the groups are defined by the interviewer who has conducted the interview. Now the grouping variable is the interviewer, and this is not directly malleable, it is not necessary to consider omitting the item that is interacting with the group variable. In this case it may be possible to change the rating behaviour of the interviewer and remove the effect of the group determiner. This was

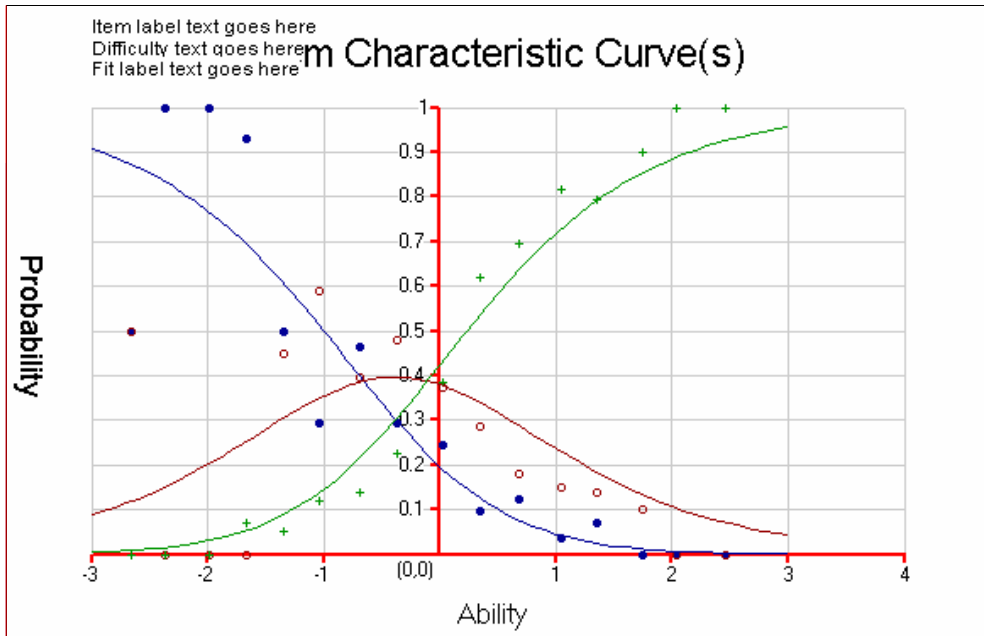
the thinking behind this analysis. If the interviewer were found to be importantly different from the benchmark interview then retraining could lead to improvement in the rating behaviour then the diff could be removed rather than removing the item, the rater or the criterion.

In this model the probability of an assigned score is determined by the ability of the candidate and the difficulty of each score point for each item. It is an extension of the simple model. The relationship between probability and the score assigned is described in a plot called the characteristic curve. For a dichotomous item, these are as in Figure 1.



Comment [g1]: Nathan, can you replace this with a better example

For the partial credit item that are as shown in Figure 2.



Comment [g2]: replace this one as well

When the data obtained for the same item over different groups of candidates are analyzed separately, the ICCs should be coincident. Small variations are tolerable because of measurement error, but in general the ICCs for different groups of candidates should not be separated. This is because of the requirement that the probability of a given score being assigned should be identical for all candidates of the same ability. When the ICCs separate for groups of candidates and do so on a systematic basis, there a secondary effect is operating and in this case we have assumed that it is the rater. The effect was that raters had difference judging patterns or stringency. A harsher judge requires a higher ability of a candidate for a specific score and a more lenient judge requires a lower score. In this case the harsher judge would generate a set of scores that would be represented by an ICC that moves to the right of the modeled curve in Figure 2 and a lenient judge would generate scores that would lead to an ICC to the left of the modeled curve. If the movement were such that the chances of selection, placement or other decision was affected by the difference in judges stringency then there is an important different function in the ICCs and some remedial action would be needed. In most cases, the item was regarded as being faulty and was described as being *biased* against specific groups. Under these circumstances the item was removed from the test. In other cases the problem was identified as being embedded in the criterion or the rating rules and these were modified. In this case

were focusing on the raters themselves and differences in stringency were identified as the basic cause of the diff. This lead us to identify raters who were systematically lenient or harsh and consequent action could be taken through retraining and adjustment of rating behaviour. Such as approach was possible using an item response modeling approach.

The approach

The study was conducted with a group of 14 interviewers each interviewing a small sample of 25 applicants. The trainer of the interviewers was an experienced language testing specialist. After conducting some 50 interviews, the trainer’s pattern of interview scores was plotted against the ‘modelled’ relationship between expected scores and assigned scores. The closeness of these patterns meant that the trainer’s scoring patterns could be used as a benchmark for assessing other interviewers. In the example below, (Item 5) it is clear that both the benchmark and interviewer’s patterns of assigned scores were very close to the modelled or theoretical pattern of expected scores given candidate proficiency.

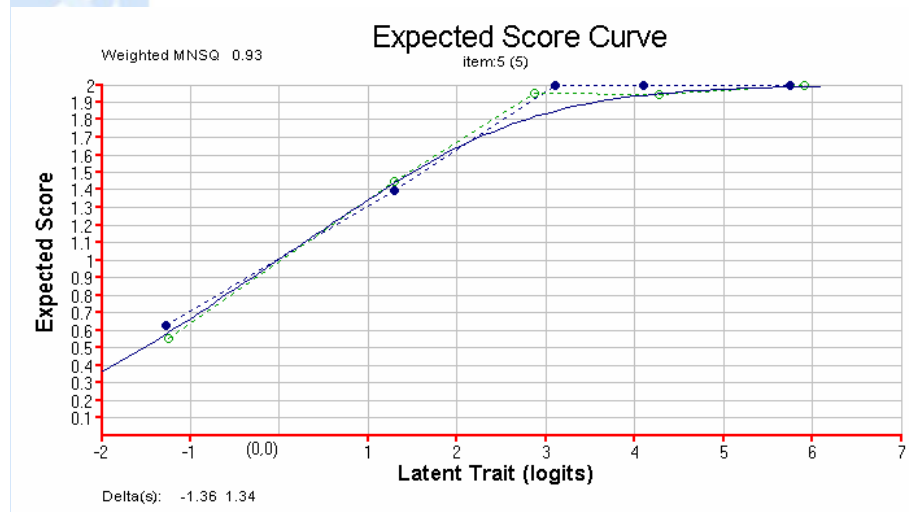
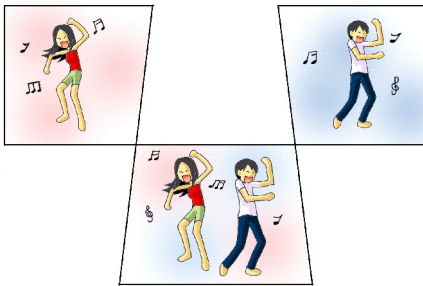


Figure 1: Modelled, benchmark and interviewer relationship between assigned score

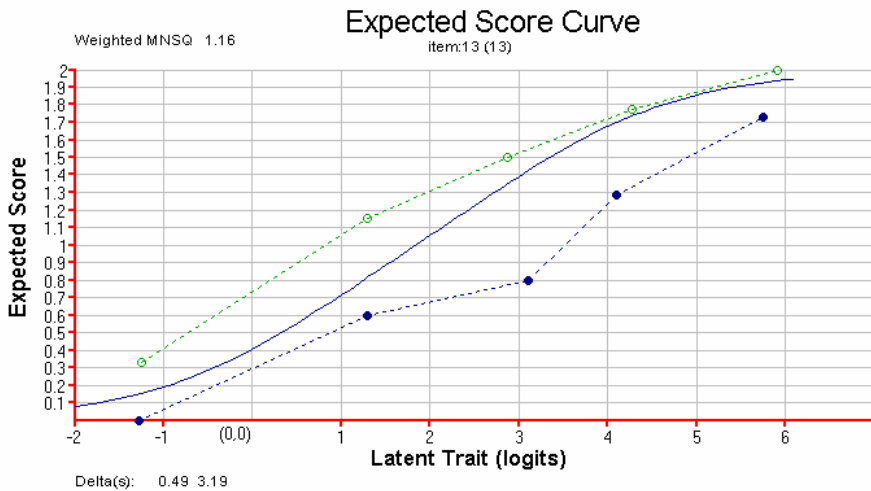
and candidate proficiency

In Figure 1 there are three curves. The smooth line represents the 'modelled' association between the expected score and the language proficiency of the candidates. The blue dashed curve represents the relationship between the interviewer's score assigned and the average language proficiency of the candidate. The green dashed curve represents the relationship between the benchmark interviewer's score assigned and the average language proficiency of the candidate.

Item 5: Describe action involving different people.
Say Stimulus



Instructions:
a. Place the picture in front of the student or indicate the room or other context as the source of people to be described.
b. Give one Say by providing three characteristics and invite the student to continue providing three further descriptions after the interviewer indicates the person(s) to be described.
Say: Look at him. He's running.... and her? ... and them? (repeat for walking and dancing)
Response Criteria
1. Neither 's/he' or 'they' is used as required.
2. One of 's/he' or 'they' is used.
3. Both 's/he' and 'they' are used as required.
All responses may include hesitation, uneven fluency and self-correction. Pronunciation may be poor, but the pronoun must be clearly understood. The pronoun may occur with or without the verb "to be", that is "s/he's", "s/he is" and "s/he", are all considered acceptable. Similarly "they're", "they are" and "they" are all considered appropriate.



Item 13 Give simple directions
Stimulus



Instructions:

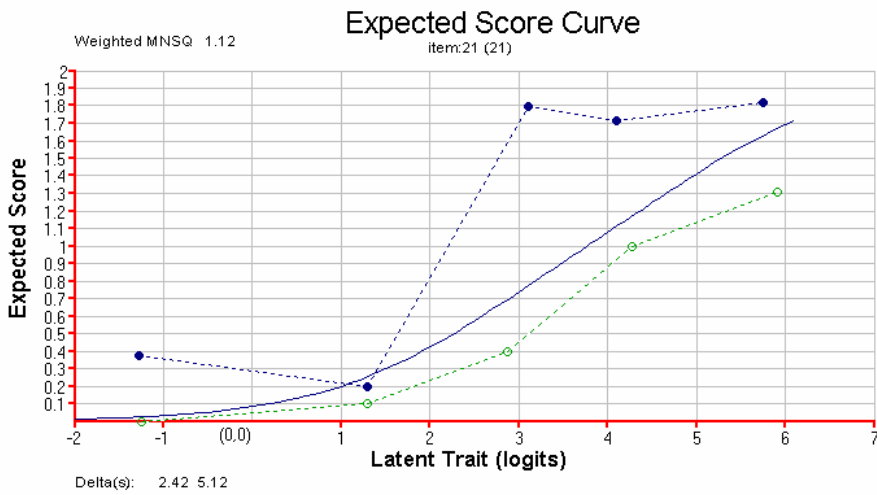
Point to the section labelled 'John's house'
 Point to the railway station.
 If the student has difficulty, break up the task by pointing out the route down Park Street, right into First Avenue and left at Station Street. But do not state the directions
 Say: I want to go from John's house to the railway station. Tell me how to get there.

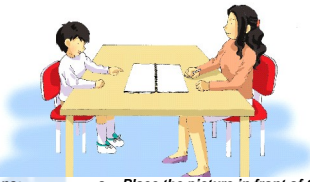
Say:

Response Criteria

1. No clear directions given OR unsuccessful direction
2. information conveyed, discounting errors. Siple prepositions are used "at", "to", "in", "on", etc...
3. Information correctly conveyed with a range of vocabulary indicating direction and purpose "opposite", "straight ahead", "along", "until" etc...

In this example Figure 2, the interviewer was consistently more stringent than either the modelled or the benchmark interviewer.





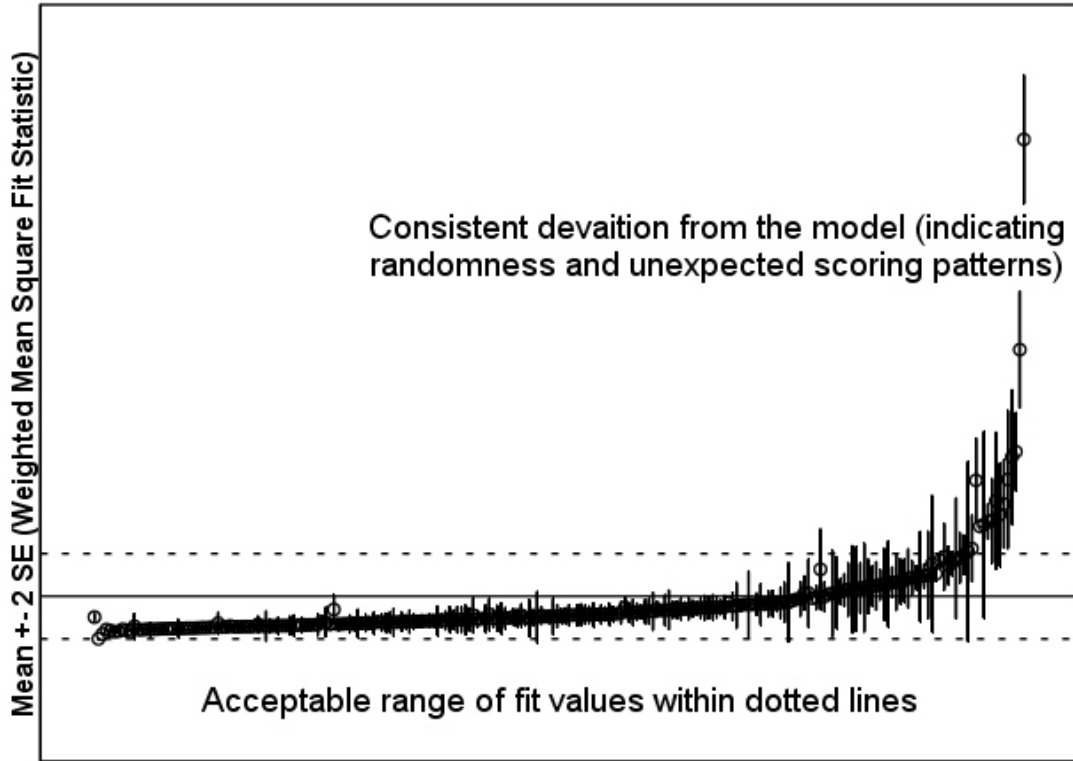
- Instructions:**
- Place the picture in front of the student, or indicate another context.
 - Explain the context of the interview as below.
- Say:**
- This is the teacher.
This is the student.
The teacher wants to know about the student.
What questions would the teacher ask the student about life and family?
What questions would the teacher ask the student about learning?
Tell me the questions and pretend you are the teacher.
- Response Criteria**
- Lacks functional competence, but can convey meaning. Incorrect sentence structure can impede meaning.
 - Errors in grammar and vocabulary interfere with communication. Articulation stress and intonation make communication possible with repetition, delivery is hesitant.
 - Sustains a series of relevant questions with a range of question forms and structures. Articulation, stress and intonation are clear and aid communication. Self correction can be present.

In this example the interviewer was displaying inconsistent judgement and was considerably more stringent than either the modelled or the benchmark curves.

Group analysis.

While these examples show that individual interviewers can be helped to improve interview judgement by being given direct feedback on specific items and even specified criteria, it was also possible to identify groups of interviewers who were not consistent with the judgement patterns of their peers. This was conducted using the fit to the Rasch model as explained above. In Hong Kong a children's version of the ITEL test was used with 1738 P1, 1208 P2 and 1536 P3 pupils randomly selected, It was administered by 615 teachers of English supervised by 89 native English Speaking Teachers. The study is reported in full by Griffin and Woods (2006). This enabled an examination of the patterns of ratings by interviewers and to identify those whose patterns of ratings did not conform with the general trend. With such large numbers of interviewers it was not possible to examine the individual stringency using the diff approach. However, even given the robustness of its analysis it did highlight groups of teachers who needed additional training in interviewing. The range of fit analyses is shown in Figure XX below. It is clear that there were a large range of interviewers who were providing possibly inconsistent ratings patterns and further training was needed for these teachers. The retraining is taking place at the time of this presentation. In part the retraining involves review of procedures, but it also required a self-examination of the data provided by each teacher.

ITEL Mean Person Fit by Rater



References

Griffin, P. and Woods K. Progress report on the Hong Kong PNET Program. Report submitted to the Hong Kong Education and Manpower Bureau. Melbourne :

Assessment Research Centre, University of Melbourne

Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer & H. Braun (Eds.), *Test Validity* (p 129-145). Hillsdale, NJ: Lawrence Erlbaum Associates.

Kelderman, H., and Macready, George B. (1990). The Use of Loglinear Models for Assessing Differential Item Functioning Across Manifest and Latent Examinee Groups. *Journal of Educational Measurement*, 27(4), 307-327.

Swaminathan, H., and Rogers, H.J. (1990). Detecting Differential Item Functioning Using Logistic Regression Procedures. *Journal of Educational Measurement*, 27, 361-370.

Wainer, H. (1993). Model-Based Standardized Measurement of an Item's Differential Impact. In P.W. Holland & H. Wainer (Eds.), *Differential Item Functioning: Theory and Practice* (p123-135) . Hillsdale, NJ:Lawrence Erlbaum.

This document was created with Win2PDF available at <http://www.daneprairie.com>.
The unregistered version of Win2PDF is for evaluation or non-commercial use only.