Investigating DIF by Statistical Modeling Of The Probability Of Endorsing An Item: Logistic Regression and Extensions Thereof

Bruno D. Zumbo
Measurement, Evaluation, & Research Methodology Program, and
Department of Statistics
University of British Columbia

Presented at the National Council for Measurement in Education meeting, April 2001. Part of the invited symposium *Differential Item Functioning: State of the Art* organized by Brian Habing, University of South Carolina.

Send correspondence to:

Bruno D. Zumbo, Ph.D.
University of British Columbia
Scarfe Building, 2125 Main Mall
Department of ECPS
(Program Area: Measurement, Evaluation, & Research Methodology)
Vancouver, B.C.
CANADA V6T 1Z4

e-mail: bruno.zumbo@ubc.ca

homepage url: http://www.educ.ubc.ca/faculty/zumbo/zumbo.html

Investigating DIF by Statistical Modeling Of The Probability Of Endorsing An Item: Logistic Regression and Extensions Thereof

Bruno D. Zumbo
Measurement, Evaluation, & Research Methodology Program, and
Department of Statistics
University of British Columbia

… an item shows DIF if individuals having the same ability, but from different groups, do not have the same probability of getting the item right (Hambleton, Swaminathan, & Rogers, 1991, p. 110)

Let us begin in the beginning …. with binary item responses.

- ❑ Uniform DIF exists when the probability of answering the item correctly is greater for one group than the other uniformly over all levels of ability: There is no interaction between ability level and group membership.
- ❑ Nonuniform DIF exists when the probability of answering the item correctly is not greater across all levels of ability for any group: There is interaction between ability level and group membership.

The logistic regression (LogR) method for detecting DIF (Swaminathan & Rogers, 1990) is based on statistical modeling of the probability of endorsing an item by group membership g, and a criterion variable *x* using logistic regression procedures. The LogR procedure allows one to model uniform and/or non-uniform DIF while using the continuous criterion variable.

With the LogR procedure for DIF one can model, for each item, the probability of observing the event *u*=1as

$$P(u=1\,|\,x,g) \;=\; \frac{\exp\{b_0+b_1x+b_2g+b_3(xg)\}}{1+\exp\{b_0+b_1x+b_2g+b_3(xg)\}}\;, \tag{1}$$

where, the response variable, *u*, is a binary variable (1=endorsing an item, 0 otherwise) denoting the item response, *g* denotes the group membership, *x* criterion variable, and *xg* the interaction. For those used to seeing “linear” models, Equation (1), which is nonlinear with respect to the odds or probabilities, can be conveniently re-expressed as linear with respect to the logits,

$$\ln\left[\frac{\pi_i}{(1-\pi_i)}\right]=b_0+b_1x+b_2g+b_3(xg), \tag{2}$$

where $\pi_i = \mathrm{P}(u=1\,|\,x,g)$. The *transformed logistic regression model for DIF* of equation (2) is sometimes called the linear probability model.

Two points are noteworthy from equations (1) and (2). First, in common applications of LogR DIF procedures *x* is the observed total scale score, *g* is an indicator variable such as

$$g = \begin{cases} 1 \text{ if the examinee is a member of Group 1} \\ 0 \text{ if the examinee is a member of Group 2,} \end{cases} \tag{3}$$

and *xg* is the product of the two explanatory variables, *g* and *x*. To avoid collinearity problems between the interaction term and the remainder of the explanatory variables in the case where there are an unbalanced number of cases in Groups 1 and 2, it is sufficient to center *x* and *g* before created the product term, *xg*. One way of centering *g* is to assign the cases in Group 1 the value $1/n_1$ and Group 2 the value $-1/n_2$, where n_1 and n_2 are the sample sizes for Groups 1 and 2, respectively.

Second, it is fruitful to conceive of logistic regression analysis as having the same goal as that of any model-building technique used in statistics so that one can apply commonly know strategies from ordinary regression to logistic regression: diagnostic measures like influence, outliers, and collinearity diagnostics, and the assessment of fit.

Let me describe the most common use of LogR wherein one uses the observed total score as the conditioning variable. For each item, the model in equation (2) is fit so that the item is first conditioned on *x*, total test score, then the presence of DIF is tested by examining the two-degree-of-freedom Chi-square test of improvement of fit of the model associated with adding *g* and the interaction term *xg* simultaneously to the model. Note that *x* and *g* were centered (as discussed above) before the product interaction term was computed. Because centering is a linear transformation it does not alter the results of the chi-square test; therefore, the results of the significance test would be the same if instead we had used 0,1 or 1,-1 indicator coding for *g*. Today, most commercial available statistical packages have a logistic regression routine available, however, in simulation studies many of us still use Judy Spray's program.

Some Extensions of Swaminathan and Rogers' Original Approach as an Indicator of Where Things Are Going in the LogR Tradition of Modeling DIF

I will provide a brief sketch of some of the extensions of the LogR DIF modeling tradition. I do so with the understanding that it will be a biased picture of my own thoughts and directions I am taking with this approach. First, however, let me highlight some points from the research literature.

1. Swaminathan and Rogers (1990) and later Swaminathan (1994) described the intimate connection between the LogR approaches for binary items and item response theory. Swaminathan (1994) writes:

   > In the model given in equation (1), if the observed *x* is replaced by the latent trait θ, we obtain the IRT formulation. Replacing the observable *x* by the unobservable θ introduces the usual complications that attend parameter estimation in item response models. Because of this relationship between the logistic regression and IRT models, and the high correlation usually found between the estimate of θ and *x* (when *x* is the total score

> on the test) we can expect … that the two procedures will agree closely. Because the IRT formulation essentially treats the score *x* as an unobservable, the logistic regression procedure is a special case of the IRT procedure. However, the logistic regression procedure does not have the estimation problems encountered in IRT and hence is a viable alternative to the IRT procedures, particularly in small samples. (p. 175)

2. Miller and Spray (1993) describe a logistic regression procedure that extends the logistic regression model described by Swaminathan and Rogers (1990). Miller and Spray also discuss a logistic discriminant function analysis (LDFA) procedure in which probabilities of group membership are predicted from item and total test scores.

3. When a test (or performance task) taps a number of abilities, a multivariate matching procedure may be most appropriate. Matching on several abilities and/or on some other background variable (e.g., an educational background variable) is clearly fairly straightforward in LogR methods.

4. Building on the notion that it is fruitful to conceive of logistic regression analysis as having the same goal as that of any model-building technique used in statistics, one can apply procedures for ordering of explanatory variables in regression analysis to address issues of explanatory variable importance leading to some measures of effect size for uniform and non-uniform DIF. Of particular interest in the context of LogR DIF procedures may be the variable ordering procedures based on partitioning an R^2 measure due to the natural (ordering) hierarchy of variables (Zumbo, 1999). One would then get a measure of the proportion of the model R^2 attributable to each term in Equations (1) and (2). Of course, this has required either: (a) an R^2-like index for LogR, or (b) the use of an ordinal logistic regression approach.

   As I describe in my Handbook, the ordinal LogR can be used with scales or tests comprised of: (a) binary items, (b) ordered polytomous items, or (c) some combination of binary and ordered polytomous items.

5. Jodoin and Gierl (in press) applied an effect size measure developed by Zumbo and Thomas (1996; also see Zumbo, 1999) for the LogR DIF procedure to investigate whether it is fruitful to use this effect size measure along with the significance test result in decision-making about DIF. First, they developed a new classification method and then applied it in a simulation study of the two-degree-of-freedom chi-squared test of DIF. Second, they tested in their simulation a suggestion by Zumbo (1999) to use single-degree-of-freedom chi-squared tests with their corresponding effect size measures. Their simulation results indicate that overall there is a great deal of promise in combining the effect measure with the significance test results in making a decision. As well, the single-degree-of-freedom tests, used along with the corresponding effect sizes, performed well.

6. Speaking of effect size measures, a natural measure when investigating the uniform DIF (with a single-degree-of-freedom test) is to compute the odds ratio for that effect. Of course, matters become far more complex with non-uniform DIF but perhaps a conditional odd-ratio at various points along the continuum of variation may work as a measure of effect size in non-uniform DIF. This may have to be prefaced with some sort of Johnson-Neyman-like procedure for the generalized linear model. The Johnson-Neyman approach would allow me to investigate what the effect of the conditioning variable for different values of the conditioning variable. It is important, at this point, to note that one can consider LogR as a case of a generalized linear model wherein we consider the stochastic structure of the data in terms of the Bernoulli and binomial distributions, and the systematic structure in terms of the logit transformation. The result is a generalized linear model (GLIM) with binomial response and logit link. One can get a great deal of mileage from the GLIM framework.

7. Finally, I am currently working on the problem of studying DIF in the context of complex survey data of the sort obtained in national population studies such as Statistics Canada's national longitudinal study of children and youth. In particular, I am working on how one would study DIF when the sample is drawn in a complex sampling structure involving, for example, clusters or two-stage sampling (a similar situation appears to occur in the NAEP context). A third example of complex data structures occurs when one examines the DIF of SAT tests. In this case, one may have a classroom, school, or district clustering effect that results in a complex population structure. In all of cases, the data are not simple random samples (SRS), the assumption that underlies all of the testing approaches currently available for DIF. The effect of not having an SRS and possibly some complex structure means that the Type I error rates of the DIF tests are most certainly significantly inflated above nominal levels. In fact, it is the complex structure of student data that has been the major impetus behind the developments in multi-level statistical approaches such as HLM. Note that simulation studies of DIF reflect a scenario where we have SRS – i.e., no classroom, school, or state level complex structures – and hence the population being sampled is not structured in any relevant manner.

   One could approach a solution to this matter from several different vantage points:
   (a) a hierarchical linear model / random regressors approach,
   (b) the analysis of complex survey data approaches taking into account the design effects, or
   (c) the approach I am favoring right now, using generalized estimating equations (GEE) as, for example, a why of approximating the underlying covariance matrix of the correlated within-cluster observations. Of course, the GEE methods can be incorporated in the LogR approach.

References

Jodoin, M. G., & Gierl, M. J. (in press). Evaluating Type I Error and Power Rates Using an Effect Size Measure with the Logistic Regression Procedure for DIF Detection. *Applied Measurement in Education.*

Miller, T. R., & Spray, J. A. (1993). Logistic discriminant function analysis for DIF identification of polytomously scored items. *Journal of Educational Measurement, 30,* 107-122.

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

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.

Zumbo, B. D. (1999). *A handbook on the theory and methods of differential item functioning (DIF): Logistic regression 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.
{freely available at url: http://www.educ.ubc.ca/faculty/zumbo/DIF/index.html}

Zumbo, B. D., & Thomas, D. R. (1996). *A measure of DIF effect size using logistic regression procedures*. Paper presented at the National Board of Medical Examiners, Philadelphia, PA.