DOCUMENT RESUME

ED 384 629 TM 023 779

AUTHOR Hanson, Bradley A.; Feinstein, Zachary S.
TITLE A Polynomial Loglinear Model for Assessing Differential Item Functioning.
PUB DATE Apr 95
NOTE 51p.; Paper presented at the Annual Meeting of the American Educational Research Association (San Francisco, CA, April 18-22, 1995).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC03 Plus Postage.
DESCRIPTORS *Equated Scores; *Item Bias; Test Format; *Test Items
IDENTIFIERS Dichotomous Variables; *Logit Analysis; *Log Linear Models; Nonequivalent Control Groups; Polytomous Items

ABSTRACT

This paper discusses loglinear models for assessing differential item functioning (DIF). Loglinear and logit models that have been suggested for studying DIF are reviewed, and loglinear formulations of the logit models are given. A polynomial loglinear model for assessing DIF is introduced. Two examples using the polynomial loglinear model for investigating DIF are discussed. One example investigates DIF for a test consisting of both dichotomous and polytomous items. Another example illustrates the use of DIF techniques in investigating whether common items are functioning differently on two forms of a test in the common item nonequivalent groups equating design. (Contains 17 references, 6 tables, and 12 figures.) (Author)

A Polynomial Loglinear Model for Assessing Differential Item Functioning

Bradley A. Hanson
American College Testing

Zachary S. Feinstein
University of Minnesota

Address correspondence to: Brad Hanson
ACT
P.O. Box 168
Iowa City, IA 52243
Ph: (319) 337-1456
email: bhanson@blue.weeg.uiowa.edu

Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco April 1995.

Abstract

This paper discusses loglinear models for assessing differential item functioning (DIF). Loglinear and logit models that have been suggested for studying DIF are reviewed, and loglinear formulations of the logit models are given. A polynomial loglinear model for assessing DIF is introduced. Two examples using the polynomial loglinear model for investigating DIF are discussed. One example investigates DIF for a test consisting of both dichotomous and polytomous items. Another example illustrates the use of DIF techniques in investigating whether common items are functioning differently on two forms of a test in the common item nonequivalent groups equating design.

There are many procedures a researcher may use to examine the validity of a test, so as to prevent bias from inadvertently affecting a sub-group of people that the test is intended for. Procedures of this type are part of the process of construct validation. One aspect of investing the validity of a test for various groups of people is the investigation of whether bias exists at the item level. Item bias is said to exist when an item is functioning differently for two or more groups of people, within the population the test is intended for. Item bias manifests itself by differential response to an item based on the group a person belongs to, when conditioned on the latent variable being measured by the test the item is a part of. The phrase *differential item functioning* (DIF) has been used to refer to this type of differential item performance. Item bias is defined conditioned on the latent variable measured by the test as there can be differences in responding to an item among groups (termed impact) that reflect legitimate differences between the groups on the latent variable measured by the test. When conditioned on the latent variable measured by the test there should be no differences between the groups in responding to the item.

This paper discusses loglinear models used for assessing DIF. The loglinear models allow investigation of DIF for dichotomously scored items (items scored correct or incorrect), or polytomously scored items (items with more than two response categories). A definition of DIF is first presented and is followed by a review of contingency table approaches that have been used to investigate DIF. A polynomial loglinear model for assessing DIF which incorporates the a numerical score given to item response categories and matching variable categories is presented. Two examples using the polynomial loglinear model for investigating DIF are given. One example investigates DIF for a test consisting of both dichotomous and polytomous items. Another example uses DIF techniques in investigating whether common items are functioning differently on two forms of a test in the common item nonequivalent groups equating design.

Definition of DIF

The data used to investigate DIF for a particular item consists of three variables: 1) an item response variable (Y), 2) a group variable (V), and 3) a matching variable (Z). It is assumed in this paper that the matching variable and item response are categorical rather than continuous (the group variable is also categorical). The data used to investigate DIF for a particular item are then contained in an $I \times J \times K$ table, where there are I categories for the for the item response, J groups, and K categories for the matching variable.

There is no DIF for the item in question if Y and V are conditionally independent given Z. Conditional independence of Y and V given Z can be expressed as

$$\Pr(Y = y, V = v \mid Z = z) = \Pr(Y = y \mid Z = z) \Pr(V = v \mid Z = z), \tag{1}$$

for all y, z, and v. Another way to express the conditional independence of Y and V given Z is

$$\Pr(Y = y \mid Z = z, V = v) = \Pr(Y = y \mid Z = z) \tag{2}$$

for all y, z, and v. The equivalence of Equations 1 and 2 is called the Fundamental Lemma of Measurement Invariance by Meredith and Millsap (1992).

True DIF is defined with the matching variable being the latent variable measured by the test the item is part of. In practice it is not possible to use this true matching variable. In this paper Z is considered to be an observable variable, in which case the condition represented by Equation 2 is referred to as observed conditional invariance (Millsap and Everson, 1993).

Let m_{ijk} be the expected count for item response category i, group j, and matching variable category k. Conditional independence of Y and V given Z is equivalent to the conditional odds ratios

$$\theta_{ij(k)} = \frac{m_{ijk}m_{i+1,j+1,k}}{m_{i+1,j,k}m_{ij+1,k}} \quad 1 \le i < I,\ 1 \le j < J \tag{3}$$

all being equal to 1 for every k. If any of the conditional odds ratios $\theta_{ij(k)}$ differs from 1 then DIF is said to exist. Uniform DIF is said to exist when some $\theta_{ij(k)}$ differ from 1 and for each i and j, $\theta_{ij(k)} = \theta_{ij(k')}$ for all $k \neq k'$. DIF that is not uniform is called nonuniform DIF.

Loglinear and Logit Models for Studying DIF

The saturated loglinear model for the three-way table of item response category by group by matching variable category is (Mellenbergh, 1982):

$$\log(m_{ijk}) = \mu + \lambda_i^Y + \lambda_j^V + \lambda_k^Z + \lambda_{ij}^{YV} + \lambda_{ik}^{YZ} + \lambda_{jk}^{VZ} + \lambda_{ijk}^{YVZ}. \tag{4}$$

One constraint is placed on each of the parameters $\lambda_i^Y, \lambda_j^V, \lambda_k^Z$ to identify the model (for example, $\lambda_1^Y = \lambda_1^V = \lambda_1^Z = 0$). Constraints are also placed on the λ_{ij}^{YV} ($I + J - 1$ constraints, for example $\lambda_{1j}^{YV} = \lambda_{i1}^{YV} = 0$ for all i, j), λ_{ik}^{YZ} ($I + K - 1$ constraints, for example $\lambda_{1k}^{YZ} = \lambda_{i1}^{YZ} = 0$ for all i, k), and λ_{jk}^{VZ} ($J + K - 1$ constraints, for example $\lambda_{1k}^{VZ} = \lambda_{j1}^{VZ} = 0$ for all j, k). There are $IJ + IK + JK - I - J - K + 1$ constraints placed in the λ_{ijk}^{YVZ}, for example $\lambda_{1jk}^{YVZ} = \lambda_{i1k}^{YVZ} = \lambda_{ij1}^{YVZ} = 0$ for all i, j, k. The model in Equation 2 has no residual degrees of freedom (the model fits any data perfectly) because it is a saturated model.

The log of the conditional odds ratios in Equation 3 for the model in Equation 4 are

$$\log\left(\frac{m_{ijk}m_{i+1,j+1,k}}{m_{i+1,j,k}m_{ij+1,k}}\right) = \lambda_{ij}^{YV} + \lambda_{i+1,j+1}^{YV} - \lambda_{i+1,j}^{YV} - \lambda_{i,j+1}^{YV} + \lambda_{i,k}^{YVZ} + \lambda_{i+1,j+1,k}^{YVZ} - \lambda_{i+1,j,k}^{YVZ} - \lambda_{i,j+1,k}^{YVZ}. \tag{5}$$

The log-odds ratios given in Equation 5 will generally differ from zero and will not be constant across levels of the matching variable category. Thus, the DIF implied by the model in Equation 5 is nonuniform DIF.

Mellenbergh (1982) identifies two nonsaturated models based on Equation 4 that are of interest in the analysis of DIF — one for uniform DIF and one for no DIF. The uniform DIF model is obtained by eliminating the λ_{ijk}^{YVZ} terms from the model of Equation 4:

$$\log(m_{ijk}) = \mu + \lambda_i^Y + \lambda_j^V + \lambda_k^Z + \lambda_{ij}^{YV} + \lambda_{ik}^{YZ} + \lambda_{jk}^{VZ}, \tag{6}$$

with the same constraints on the parameters as were indicated for the model in Equation 4. The log of the odds ratios in Equation 3 for the model in Equation 6 are

$$\log\left(\frac{m_{ijk}m_{i+1,j+1,k}}{m_{i+1,j,k}m_{ij+1,k}}\right) = \lambda_{ij}^{YV} + \lambda_{i+1,j+1}^{YV} - \lambda_{i+1,j}^{YV} - \lambda_{i,j+1}^{YV}. \tag{7}$$

The log-odds in Equation 6 will in general differ from 1 but do not differ across levels of the matching variable. Thus, the model given by Equation 6 implies uniform DIF.

The no DIF model presented by Mellenbergh (1982) is obtained by eliminating the λ_{ij}^{YV} terms from the model in Equation 6:

$$\log(m_{ijk}) = \mu + \lambda_i^Y + \lambda_j^V + \lambda_k^Z + \lambda_{ik}^{YZ} + \lambda_{jk}^{VZ} . \tag{8}$$

The log of the odds ratios in Equation 3 for the model in Equation 8 will all be zero. Thus, the model in Equation 8 implies no DIF for the item.

To use the models in Equations 4, 6 and 8, Mellenbergh (1982) suggests that first the model in Equation 6 be fit to the data. If this model does not fit the data (based on Pearson or likelihood ratio chi-squared statistics) this implies nonuniform DIF for the item. If the model in Equation 6 does fit the data then the model in Equation 8 is fit to the data. If the difference in the likelihood ratio chi-squared statistics for the models in Equations 8 and 6 is significant this implies the item exhibits uniform DIF. If the difference in the chi-squared statistics for the models in Equations 8 and 6 is not significant then the item does not exhibit DIF.

Logit Models

Mellenbergh (1982) notes that for dichotomous items logit models equivalent to the loglinear models in Equations 4, 6 and 8 for the purposes of studying DIF can be used. In the logit models the response variable is $\log(m_{1jk}/m_{2jk})$, where there are only two categories of item response.

In the case in which there are numeric scores associated with the matching variable categories and/or item response categories this information can be used to create more parsimonious logit (and loglinear) models for studying DIF. Let the scores associated with the item response categories be $r_1, r_2, \ldots, r_I$, and let the score associated with matching variable categories be $s_1, s_2, \ldots, s_K$. It is assumed the categories are arranged such that $r_1 \leq r_2, \ldots, \leq r_I$ and $s_1 \leq s_2, \ldots, \leq s_K$.

In the case of a dichotomous item response Swaminathan and Rogers (1990) present logit models where linear functions of the matching variable score are substituted for the nominal matching variable effects in the logit models presented by Mellenbergh (1982). This allows for a nonsaturated logit model for nonuniform DIF (Mellenbergh's logit model for nonuniform DIF is a saturated model). The model presented by Swaminathan and Rogers (1990) can be written as

$$\log\left(\frac{m_{1jk}}{m_{2jk}}\right) = \alpha_0 + \lambda_j^V + \alpha_1 s_k + \alpha_{2j} s_k , \tag{9}$$

where there is one constraint put on the λ_j^V (for example, $\lambda_1^V = 0$), and one constraint is put on the α_{2j} (for example, $\alpha_{21} = 0$). The logit model in Equation 9 is equivalent to the following loglinear model (Agresti, 1990, pages 152-153)

$$\log(m_{ijk}) = \mu + \lambda_i^Y + \lambda_j^V + \lambda_k^Z + \lambda_{ij}^{YV} + \lambda_{jk}^{VZ} + \beta_i s_k + \gamma_{ij} s_k . \tag{10}$$

The same constraints are put on the parameters λ_i^Y, λ_j^V, λ_k^Z, λ_{jk}^{VZ}, and λ_{ij}^{YV} as were put on the corresponding parameters for the model in Equation 4. One constraint is placed on the β_i (for example, $\beta_1 = 0$) and $(I-1)(J-1)$ constraints are placed on the γ_{ij} (for example, $\gamma_{1j} = \gamma_{i1} = 0$ for all i, j). For the loglinear model in Equation 10, unlike the logit model in Equation 9, it is possible that the number of item response categories could be greater than 2. The log of the odds ratios in Equation 3 for the model in Equation 10 are given by

$$\log\left(\frac{m_{ijk} m_{i+1,j+1,k}}{m_{i+1,j,k} m_{i,j+1,k}}\right) = \lambda_{ij}^{YV} + \lambda_{i+1,j+1}^{YV} - \lambda_{i+1,j}^{YV} - \lambda_{i,j+1}^{YV}$$
$$+ (\gamma_{ij} + \gamma_{i+1,j+1} - \gamma_{i+1,j} - \gamma_{i,j+1}) s_k . \tag{11}$$

The log of the odds ratios in Equation 11 are linear functions of the matching variable score and therefore represent nonuniform DIF.

Eliminating the γ_{ij} terms from the model in Equation 10 gives

$$\log(m_{ijk}) = \mu + \lambda_i^Y + \lambda_j^V + \lambda_k^Z + \lambda_{ij}^{YV} + \lambda_{jk}^{VZ} + \beta_i s_k \,. \tag{12}$$

The log of the odds ratios in Equation 3 for the model in Equation 12 are

$$\log\left(\frac{m_{ijk}m_{i+1,j+1,k}}{m_{i+1,j,k}m_{i,j+1,k}}\right) = \lambda_{ij}^{YV} + \lambda_{i+1,j+1}^{YV} - \lambda_{i+1,j}^{YV} - \lambda_{i,j+1}^{YV} \,. \tag{13}$$

Equation 13 is in general different from zero but constant for all values of the matching variable score. Consequently, the model in Equation 12 represents uniform DIF. The log of the odds ratios in Equations 13 and 7 are identical since the only difference between the models in Equations 12 and 6 are the interaction terms involving the item response and matching variable which cancel out when computing the odds ratio.

Eliminating the λ_{ij}^{YV} from the model in Equation 12 gives

$$\log(m_{ijk}) = \mu + \lambda_i^Y + \lambda_j^V + \lambda_k^Z + \lambda_{jk}^{VZ} + \beta_i s_k \,. \tag{14}$$

The log of the odds ratios in Equation 3 for the model in Equation 14 are all zero. Consequently, the model in Equation 14 represents no DIF.

Comparing the fits of the models in Equations 10 and 12 gives a test for nonuniform DIF, and comparing the fits of the models in Equations 12 and 14 gives a test of uniform DIF.

For the case in which there are two groups but more than two item response categories Miller and Spray (1993) suggest using a logit model with group as the response variable. Their model can be written as

$$\log\left(\frac{m_{i1k}}{m_{i2k}}\right) = \alpha_0 + \alpha_1 s_k + \alpha_2 r_i + \alpha_3 s_k r_i \,. \tag{15}$$

The logit model in Equation 15 can be written as the following loglinear model

$$\log(m_{ijk}) = \mu + \lambda_i^Y + \lambda_j^V + \lambda_k^Z + \lambda_{ik}^{YZ} + \beta_{1j} s_k + \beta_{2j} r_i + \gamma_j s_k r_i \,. \tag{16}$$

The same constraints are put on the parameters λ_i^Y, λ_j^V, λ_k^Z, and λ_{ik}^{YZ} as were put on the corresponding parameters for the model in Equation 4. One constraint is put on each of the parameters β_{1j}, β_{2j} and γ_j (for example, $\beta_{11} = \beta_{21} = \gamma_1 = 0$). The log of the odds ratios in Equation 3 for the model in Equation 16 are

$$\log\left(\frac{m_{ijk}m_{i+1,j+1,k}}{m_{i+1,j,k}m_{i,j+1,k}}\right) = \beta_{2,j+1} - \beta_{2j} + (\gamma_{j+1} - \gamma_j)s_k \,. \tag{17}$$

The log-odds ratios in Equation 17 will in general be different from zero and are a linear function of the matching variable score. Consequently, nonuniform DIF is implied by the model in Equation 16.

Eliminating the terms involving γ_j from the model in Equation 16 gives

$$\log(m_{ijk}) = \mu + \lambda_i^Y + \lambda_j^V + \lambda_k^Z + \lambda_{ij}^{YV} + \beta_{1j}s_k + \beta_{2j}r_i \,. \tag{18}$$

The log of the odds ratios in Equation 3 for the model in Equation 18 are

$$\log\left(\frac{m_{ijk}m_{i+1,j+1,k}}{m_{i+1,j,k}m_{i,j+1,k}}\right) = \beta_{2,j+1} - \beta_{2j} \,. \tag{19}$$

The log-odds ratios in Equation 19 will in general differ from zero, but do not vary with the matching variable score. Consequently, uniform DIF is implied by the model in Equation 18.

Eliminating the β_{2j} from Equation 18 gives

$$\log(m_{ijk}) = \mu + \lambda_i^Y + \lambda_j^V + \lambda_k^Z + \lambda_{ij}^{YV} + \beta_{1j}s_k \,. \tag{20}$$

The log of the odds ratios in Equation 3 for the model in Equation 20 will all be zero. Consequently, no DIF is implied by Equation 20.

Comparing the fits of the models in Equations 16 and 18 gives a test for nonuniform DIF, and comparing the fits of the models in Equations 18 and 20 gives a test of uniform DIF.

An advantage of using the logit form of the models in Equations 9 and 15 as opposed to the loglinear form of these models is that there are far fewer parameters to estimate in the logit formulation. A possible advantage of using the loglinear formulation of the models as opposed to the logit formulation is that the loglinear models can be generalized to deal with more than 2 item response categories ($J > 2$) and more than 2 groups ($K > 2$) without complications of having to deal with a polytomous dependent variable (Equation 9 cannot model more than two item response categories, and Equation 15 cannot model more than two groups).

The next section presents a loglinear model in which the scores on the item responses and matching variable are used in a way that results in far fewer model parameters than for the loglinear forms of the logit models, or the loglinear models presented in Equations 4, 6 and 8.

A Polynomial Loglinear Model for Studying DIF

Loglinear models with polynomial terms involving test and item scores (polynomial loglinear models) have been used in several measurement applications. For example, smoothing of test score distributions (Holland and Thayer, 1987; Kolen, 1991), equating (Rosenbaum and Thayer, 1987; Hanson, 1991; Livingston, 1993; Little and Rubin, 1994), and testing for differences in score distributions among groups (Hanson, 1992). This section presents a model for the three-way table of item response, group, and matching variable used to investigate DIF that is analogous to polynomial loglinear models previously used in the literature.

In the loglinear models in Equations 4, 6, and 8 the item response variable and matching variable are treated as nominal. When there are scores associated with the item response categories and the matching variable categories the following loglinear model can be used

$$\log(m_{ijk}) = \mu + \lambda_j^V + \sum_{g=1}^{d_1}\beta_{1gj}s_k^g + \sum_{h=1}^{d_2}\beta_{2hj}r_i^h + \sum_{g=1}^{d_1}\sum_{h=1}^{d_2}\gamma_{ghj}s_k^g r_i^h \,, \tag{21}$$

where $d_1 < K$, $d_2 < I$. As in Equation 4 a constraint is put on the λ_j^V. There are no constrains put on the β parameters. A subset of the γ_{ghj} are assumed to be nonzero, and the rest are assumed to be zero. If it is assumed $\gamma_{g^*h^*j^*} \neq 0$ for particular values g^*, h^* and j^* then it is also assumed that $\gamma_{g^*h^*j'} \neq 0$ for all $j' \neq j^*$. Consequently, the number of $\gamma_{ghj} \neq 0$ is Jd_3 for some positive integer d_3. The value of d_3 is equal to the number of the $d_1 \times d_2$ possible γ_{ghj} in each group that are specified to be nonzero. The value of d_3 is not directly related to the values of d_1 and d_2, for example, d_3 is not the sum of d_1 and d_2. Note that the models in Equations 10 and 16 are not special cases of the model in Equation 21.

The log of the conditional odds ratios in Equation 3 for the model in Equation 21 is

$$\log\left(\frac{m_{ijk}m_{i+1,j+1,k}}{m_{i+1,j,k}m_{i,j+1,k}}\right) = \sum_{h=1}^{d_2}(\beta_{2,h,j+1} - \beta_{2,hj})\,[r_{i+1}^h - r_i^h] + \sum_{g=1}^{d_1}\sum_{h=1}^{d_2}(\gamma_{gh,k+1} - \gamma_{ghk})\,[r_{i+1}^h - r_i^h]\,s_k^g\,. \quad (22)$$

Equation 22 represents nonuniform DIF. The DIF given in Equation 22 is constrained relative to the DIF given by the saturated loglinear model (Equation 4). The model in Equation 21 is a nonsaturated loglinear model that allows for nonuniform DIF. Comparing Equation 22 to Equations 11 and 17 it is seen that the loglinear model in Equation 21 allows for more complicated forms of DIF than the models in Equations 10 and 16. In particular, the model in Equation 21 allows DIF which varies across adjacent item response categories.

The constrained version of the model given in Equation 21 which implies uniform DIF is

$$\log(m_{ijk}) = \mu + \lambda_j^V + \sum_{g=1}^{d_1}\beta_{1gj}s_k^g + \sum_{h=1}^{d_2}\beta_{2hj}r_i^h + \sum_{g=1}^{d_1}\sum_{h=1}^{d_2}\gamma_{gh}s_k^g r_i^h\,. \quad (23)$$

The model in Equation 23 differs from the model in Equation 22 by not having the γ_{gh} parameters differ for the different groups. The difference in the number of parameters between the models in Equations 23 and 21 is $d_3(J-1)$. The log of the conditional odds ratios in Equation 3 for the model in Equation 23 is

$$\log\left(\frac{m_{ijk}m_{i+1,j+1,k}}{m_{i+1,j,k}m_{i,j+1,k}}\right) = \sum_{h=1}^{d_2}(\beta_{2,h,j+1} - \beta_{2hj})\,[r_{i+1}^h - r_i^h] \quad (24)$$

The constrained version of the model given in Equation 23 which implies no DIF is

$$\log(m_{ijk}) = \mu + \lambda_j^V + \sum_{g=1}^{d_1}\beta_{1gj}s_k^g + \sum_{h=1}^{d_2}\beta_{2h}r_i^h + \sum_{g=1}^{d_1}\sum_{h=1}^{d_2}\gamma_{gh}s_k^g r_i^h\,, \quad (25)$$

The model in Equation 25 differs from the model in Equation 23 by not having the β_{2h} parameters differ for the different groups. The difference in the number of parameters for the models given in Equations 25 and 23 is $d_2(J-1)$. The log of the odds ratios in Equation 3 for the model in Equation 25 are all zero.

The likelihood ratio chi-squared statistics for the models in Equations 21 and 23 can be used to test for nonuniform DIF. Under the hypothesis that the model in Equation 23 holds, the difference in the likelihood ratio chi-squared statistics between models 21 and 23 is asymptotically distributed as a chi-square random variable with $d_3(J-1)$ degrees of freedom. For a level of significance p, the hypothesis that the model in Equation 23 holds (uniform DIF) versus the alternative hypothesis that the model in Equation 21 holds (nonuniform DIF) is rejected if the difference in the likelihood ratio chi-square statistics of the models in Equations 21 and 23 is greater then the upper p percentage point for the chi-square distribution with $d_3(J-1)$ degrees of freedom.

To test for uniform DIF the likelihood ratio chi-squared statistics for the models in Equations 23 and 25 can be used. Under the hypothesis that the model in Equation 25 holds, the difference in the likelihood ratio chi-squared statistics between models 23 and 25 is asymptotically distributed as a chi-square random variable with $d_2(J-1)$ degrees of freedom. For a level of significance p, the hypothesis that the model in Equation 25 holds (no DIF) versus the alternative hypothesis that the model in Equation 23 holds (uniform DIF) is rejected if the difference in the likelihood ratio chi-square statistics of the models in Equations 23 and 25 is greater then the upper p percentage point for the chi-square distribution with $d_2(J-1)$ degrees of freedom.

Choosing a Model

Using the models in Equations 21, 23 and 25 involves choosing values for d_1 and d_2, and choosing which of the γ_{ghj} to make nonzero. The values of d_1, d_2, and which γ_{ghj} to make nonzero are chosen based on the model in Equation 21, and are used for the models in Equations 23 and 25 in testing for uniform and nonuniform DIF.

A model selection procedure presented by Haberman (1974) can be used for choosing a model in the form of Equation 21 from a set of possible models (different values of d_1, d_2 and nonzero γ_{ghj}). To apply Haberman's (1974) procedure it is assumed that a set of q models have been identified $(M_1, M_2, \ldots, M_q)$ where model M_{i-1} is nested within model M_i, $i = 2, \ldots, q$ (M_1 is the simplest model, and M_q is the most complex model). If G_i^2 is the likelihood ratio chi-square statistic for model M_i then for $i = 2, \ldots, q$, $G_{i-1}^2 - G_i^2$ is the likelihood ratio statistic for testing the null hypothesis H_{i-1} versus the alternative hypothesis H_i, where H_i is the hypothesis that model M_i holds. If the hypothesis H_{i^*} is true then the statistics $G_{i-1}^2 - G_i^2$ for $i = q, q-1, \ldots, i^*+1$ are asymptotically independent and have chi-square distributions with w_i degrees of freedom, where w_i is equal to the difference in the number of parameters of models M_i and M_{i-1}. For a level of significance p, with $p^* = 1-(1-p)^{1/(q-1)}$, the probability that $G_{i-1}^2 - G_i^2, i = q, q-1, \ldots, i^*+1$ exceeds C, the upper p^* percentage point for the chi-square distribution with w_i degrees of freedom is asymptotically no greater than p. A simultaneous test of the hypothesis $H_i, i = q-1, q-2, \ldots, 1$, is to reject all hypotheses H_i such that $i < i'$, where i' is the largest i such that $G_{i-1}^2 - G_i^2 > C$. With a specified value of p, this hypothesis testing procedure would allow one to eliminate from consideration models $M_i, i < i'$. It gives no guidance for choosing from among the models $M_i, i \geq i'$, although typically model $M_{i'}$ (the simplest model) is chosen. Smaller values of p would would make it harder to reject the null hypothesis of the simpler model and therefore favor the selection of simpler models.

The selection procedure of Haberman (1974) requires that the models being considered form a nested sequence. Especially in the case of non-dichotomous items it is possible that the set of models under consideration do not form a nested sequence. In that case the Haberman model selection procedure is not directly applicable. A series of model comparisons could be performed, but the tests would no longer be independent and the error rate given by the Haberman procedure will no longer be accurate. The first example presented in the next section provides an example of using a modification of the Haberman procedure to select a model for polytomous items.

In applied settings it may not be realistic to use a model selection procedure for each item. A more realistic procedure may be to select a common model for all items with a specific number of score categories, perhaps based on past experience. This procedure is used in the second of the following examples.

Examples

Two examples of applying the polynomial loglinear model are presented in this section. First, the polynomial loglinear model is applied to the 27 item data set analyzed in Miller and Spray (1993). The results from the polynomial loglinear model are compared to results using the logit model of Miller and Spray (1993). The second example consists of applying the polynomial loglinear model to investigate DIF for common items in a common-item nonequivalent groups equating design. The goal is to examine if any of the common items function differently on the two test dates (the common items are embedded within different test forms on the two test dates).

Matching Variable

For both examples the matching variable will be a test score consisting of the sum of the item scores. The issue discussed in this section is whether to use as the matching variable the sum of the item scores including the studied item, or the sum of the item scores excluding the studied item.

Several authors have used theoretical justifications to conclude that a matching variable that is the sum of item scores should include the studied item (Holland and Thayer, 1988; Zwick, 1990; Meredith and Millsap, 1992). If there is a latent variable under which local independence holds for the item scores then a test score which excludes the studied item score will be conditionally independent of the studied item score given the latent variable. Under this condition and some other fairly mild conditions, Meredith and Millsap (1992) show that DIF will appear when the test score excluding the studied item score is used as a matching variable even if there is no DIF in either the studied item score or the test score when the latent variable is used as the matching variable. Under these conditions even though there is no DIF when using the latent variable as the matching variable, DIF will be observed when the test score excluding the studied item score is used as the matching variable. It is only under very special conditions that using the test score including the studied item score will alleviate this problem (e.g., the Rasch model holds for the item responses, Holland and Thayer, 1988). Consequently, *theoretical analysis* suggests the problem of DIF being detected when using an observed matching variable when no DIF exists using the ideal latent matching variable will occur in many practical situations whether or not the test score used for the observed matching variable includes or excludes the studied item score.

In the cases considered in this paper the item category scores for all items are $0, 1, \ldots, I - 1$, where there I item response categories. Let m_{ijk} be the expected count corresponding to item response category i, group j, and matching variable category k, where the matching variable is the total test score excluding the score for the studied item. Let m^*_{ijk} be the expected counts in the three-way table where the matching variable is the total test score including the score for the studied item. If there are I item response categories, J groups, and K score categories for the test score excluding the studied item, then the table containing the expected counts m_{ijk} has $I \times J \times K$ cells and the table containing the expected counts m^*_{ijk} has $I \times J \times (K + I - 1)$ cells. The expected counts m^*_{ijk} can be written in terms of the expected counts m_{ijk} as

$$m^*_{ijk} = \begin{cases} m_{i,j,k-i+1} & i \leq k \leq K+i-1 \\ 0 & k < i, k > K+i-1 \, . \end{cases} \tag{26}$$

For example, consider group 1 and item response category 2. Assuming item response categories are ordered by the category scores, then item response category 2 corresponds to an item score of 1. From Equation 26, $m^*_{21k} = m_{2,1,k-1}$ for $2 \leq k \leq K+1$. This is because any examinee who obtained a score of 1 on the item would have a test score including the item that was one greater than their test score excluding the item. For $k = 1$, Equation 26 gives $m^*_{211} = 0$ since if an examinee obtained a score of 1 on the item, the test score including this item could not be zero. In addition, for $k > K+1$ Equation 26 gives $m^*_{21k} = 0$. This is because if an examinee obtained a score of 1 on the item, their maximum test score including the item corresponds to matching variable score category $K+1$ (one more than the maximum test score excluding the item).

Equation 26 shows that the expected counts in the table corresponding to the test score including the studied item can be written in terms of the expected counts in the table corresponding to test score excluding the studied item. Even though there are $IJ(I-1)$ more cells in the table corresponding to the test score including the studied item that table will have $IJ(I-1)$ cells with structural zeros. Including the studied item score in the test score creates a table with more cells but no more information. A loglinear model fit to the table corresponding to the test score excluding the studied item would give the same results as a loglinear model fit to the table corresponding to the test score including the studied item as long as the structural zeros in the table were taken into account when fitting the model. Different results would be obtained if the model fit to the table corresponding to the test score including the studied item allowed all the cells in the table to have non-zero expected counts (which would result in non-zero fitted counts for cells in which the fitted count by definition should be zero).

Consequently, in the present setting the estimated counts and model fits would be the same whether the studied item is included in the test score or not (as long as structural zeros are preserved when including the item score in the test score). The matching variable used in the examples is the test score *excluding* the item score.

Example 1

The first example uses the same data used for the example in Miller and Spray (1993). The data consists of responses of 1976 examinees to a 27 item experimental mathematics performance test. The test consisted of 12 multiple-choice items (items 1 through 12), 9 gridded-response items (items 13 through 21), and 6 open-ended items (items 22 through 27). The multiple-choice and gridded-response items were scored dichotomously (one for a correct response and zero for an incorrect response). The scores on the open-ended items were $0, 1, 2, \ldots, k$, where $k = 3, 3, 4, 4, 5, 6$ for items 22 through 27, respectively. DIF was investigated for males versus females. There were 1005 male and 971 female examinees in the data set. One male examinee included in the data analyzed by Miller and Spray (1993) was dropped from the analyses reported here because all of his responses to the polytomous items were missing.

The first step in fitting the loglinear models given in Equations 21, 23, and 25 is determining the number of parameters to use in the models (the values of d_1, d_2, and d_3). A modified version of the Haberman procedure described above is used to select values of d_1, d_2 and d_3. Models are considered with values of d_1 ranging from 1 to 6, values of d_2 ranging from 1 to the maximum score on the item ($I - 1$), and d_3 ranging from 1 to 5. The five interaction parameters considered were γ_{11}, γ_{12}, γ_{21}, γ_{13}, and γ_{31}. A model with $d_3 = l$ would include only the first l of these interaction parameters. For example, if $d_3 = 1$ then the only interaction parameter in the model would be γ_{11}. If $d_3 = 3$, then the three interaction parameters in the model would be γ_{11}, γ_{12}, and γ_{21}.

For the dichotomously scored items (items 1 through 21) the only possible value of d_2 is 1, and d_3 was set equal 1. Consequently, for these items choosing a model involves choosing a value of d_1. For the dichotomous items the Haberman procedure was applied for the sequence of models corresponding to $d_1 = 1, 2, \ldots, 6$. The overall level of significance was chosen to be .01, so the value of p^* used for each individual test of the nested models was $1 - (1 - .01)^{1/5} = .002$.

For the polytomous items (items 22 through 27) values must be chosen for d_1, d_2 and d_3 rather than for just d_1. Instead than specifying one sequence of nested models, a sequence of nested models was specified separately for d_1, d_2, and d_3. The Haberman model selection procedure was applied three times — once for d_3, once for d_2,and once for d_1. An error rate of .003 was chosen for each of the three separate Haberman procedures resulting in in an overall error rate of at most .009 (by the Bonferroni inequality) for the three procedures taken as a whole. When selecting d_3, d_1 and d_2 were set equal to their maximum values (6 for d_1, and $I - 1$ for d_2). The Haberman procedure was applied to a sequence of models given by $d_3 = 1, 2, \ldots, 5$. The overall level of significance chosen was .003, so the value of p^* used for each individual test of the nested models was $1 - (1 - .003)^{1/4} = .00075$.

When selecting d_2, d_1 was set equal to 6 and d_3 was set equal to the value determined in the first step. The Haberman procedure was applied to a sequence of models given by $d_2 = 1, 2, \ldots, I - 1$. For an overall level of significance of .003, the value of p^* used for each of the individual tests of the nested models was $1 - (1 - .003)^{1/(I-2)}$.

When selecting d_1, the values of d_2 and d_3 were set equal to the values chosen in the previous steps. The Haberman procedure was applied to a sequence of models given by $d_1 = 1, 2, \ldots, 6$. For an overall level of significance of .003, the value of p^* used for each of the individual tests of the nested models was $1 - (1 - .003)^{1/5} = .0006$.

Examples of applying the model selection procedure to items 7 and 23 are presented in Table 1. The top part of Table 1 gives results for item 7. A nested sequence of six models were compared. Chi-square statistics for comparing adjacent models and their degrees of freedom and p-values are presented in the last three columns. The first two models to be compared are those given in the first two rows. For both these models $d_2 = 1$ and $d_3 = 1$. The model in the first row has $d_1 = 6$ and the model in the second row has $d_1 = 5$. The chi-square statistic for testing the null hypothesis of the model with $d_1 = 5$ against the alternative hypothesis of the model with $d_1 = 6$ is given as .544. The value of p^* chosen for each of the tests of consecutive models is .002. Consequently, for the first test ($d_1 = 5$ versus $d_1 = 6$) the null hypothesis of the simpler model is not rejected. The first test that is significant at the .002 level is the test for $d_1 = 2$ versus $d_1 = 3$. Consequently, the model selected is $d_1 = 3$ (this is indicated in the table by the value of p^* being next to the model with $d_1 = 3$).

For item 23 separate selection procedures were used for d_3, d_2, and d_1. For item 23 the first five lines correspond to models with five different values of d_3. For these models d_1 and d_2 were fixed at their maximum values of 6 and 3, respectively. The level of significance chosen for the tests of consecutive models was .00075. The first model comparison was for $d_3 = 4$ versus $d_3 = 5$. The chi-square statistic for this test is 21.34 with 2 degrees of freedom which is significant at the

.00075 level. Consequently, the simpler model with $d_3 = 4$ is rejected, and the value of $d_3 = 5$ is chosen. Next, three models corresponding to three values of d_2 are compared (with $d_1 = 6$ and $d_3 = 5$). In this case a value of $d_3 = 3$ is selected. Finally, there are six models corresponding to $d_1 = 6, 5, \ldots, 1$ (with $d_2 = 3$ and $d_3 = 5$). A value of $d_1 = 4$ is chosen. Thus, for item 23 the model with $d_1 = 4$, $d_2 = 3$, and $d_3 = 5$ is used to test for uniform and nonuniform DIF.

For all dichotomous items, except items 7 and 13, d_1 was selected to equal 4, while for dichotomous items 7 and 13 d_3 was selected to be 3. The models chosen for the polytomous items are given in Table 2. For all the polytomous items, $d_1 = 4$ and $d_2 = I - 1$ (the maximum score on the item). Varying numbers of interactions terms were chosen for the polytomous items. For example, for polytomous item 24 only one interaction parameter between item response and score level was chosen. For polytomous item 27, four interaction parameters were chosen.

The values of d_1, d_2, and d_3 chosen were used in fitting the models in Equations 21, 23, and 25 for each item. Likelihood ratio chi-square statistics for testing for uniform and nonuniform DIF were computed. When reporting results, three levels of significance are used — 0.05, 0.01 and 0.05 / 27 = 0.00185 (a Bonferroni adjustment).

Significance levels for tests of uniform and nonuniform DIF are shown in Table 3 for all items. For uniform DIF, thirteen items reached the .05 level of significance , ten items reached the .01 level of significance, and eight items reached the .00185 level of significance. For non-uniform DIF, two items shows significant nonuniform DIF at the .05 level, and one of these items (item 15) did not show significant uniform DIF. For Item 15 nonuniform DIF was indicated, but the bias was balanced to cancel the effects against each group out and this item exhibited no uniform DIF.

The logistic discriminant function analysis (LDFA) results using the models in Equations 16, 18 and 20 are presented in Table 4. The results in Table 4 differ slightly from the results in Table 4 of Miller and Spray (1993) because the analysis reported here used a matching variable that did not include the studied item, whereas Miller and Spray (1993) used a matching variable that did include the studied item.

There is little difference between the LDFA and polynomial loglinear models in terms of the tests for uniform DIF. For the polynomial loglinear model the test for nonuniform DIF was significant at the .05 level of significance for only items 15 and 26. For the LDFA model the test for nonuniform DIF was significant for ten items at the .05 level of significance, for five items at the .01 level of significance , and for two items at the .00185 level of significance. For this data the LDFA model indicated more nonuniform DIF than the polynomial loglinear model.

The log of the odds ratios in Equation 3 for the observed counts and fitted counts for the polynomial loglinear and LDFA nonuniform DIF models were graphed and compared to explore the DIF trends across score levels. Graphs were created for dichotomous items in which significant nonuniform DIF was indicated for the LDFA model but not for the polynomial log-linear model.

The graph of the log-odds as a function of test score (excluding the studied item) is presented in Figure 1 for item 16. In all figures group 0 is males and group 1 is females.

As can be seen from the graph in Figure 1, the observed log-odds lie primarily above the line of no DIF. The line of no DIF is a horizontal line perpendicular to the y-axis (or log-odds ratio axis) at 0.0. If an item had no DIF, the observed data would approximate this line. The observed data shows scatter, but it does not appear to have any trend. As can also be seen from the graph, the fitted log-odds ratios for the LDFA model have more slope than the fitted log-odds ratios for the polynomial loglinear model. The polynomial loglinear model has very little slope, and significant nonuniform DIF was not indicated for this model. For this item the polynomial loglinear model appears to fit the observed data more closely than the LDFA model.

Log-odds plots for items 4, 5, 6, 8, and 20 are presented in Figures 2 through 6, respectively. Like item 16, significant nonuniform DIF was indicated by the LDFA model for these items, but not by the polynomial loglinear model. For all items except item 8 there does not appear to be much of a trend in the log-odds ratios, which is more consistent with the fitted odds-ratios for the polynomial loglinear model. For item 8 there may be some slight trend in the observed log-odds ratios, which is more consistent with the fitted odds-ratios for the LDFA model.

There was only one item, item 15, for which the test for nonuniform DIF was significant for the polynomial loglinear model but not for the LDFA model. For one other item, item 7, the test for nonuniform DIF was very near to being significant for the loglinear polynomial model but not for the LDFA model. Graphs of the odds ratios for items 15 and 7 are given in Figures 7 and 8.

The results in Figures 7 and 8 are similar. The fitted log-odds ratios for the polynomial loglinear model show a positive slope. There is only a very slight trend in the fitted log-odds ratios for the LDFA model, and this trend is in the opposite direction of the trend in the fitted log-odds ratios for the polynomial loglinear model. To the extent there is any trend in the observed log-odds ratios, that trend appears to be in the same direction as the trend of the fitted log-odds ratios for the polynomial loglinear model.

The test for nonuniform DIF was significant for the last 5 polytomous items using the LDFA model, whereas only for item 26 was the test for nonuniform DIF significant for the polynomial loglinear model. For the polytomous items there would be multiple log-odds ratio plots (one plot for each pair of adjacent item response categories) for each item analogous to the single plots for the dichotomous items given in Figures 1 through 8. Because of the sparseness of the data it is not practical to plot the the multiple observed log-odds as a function of matching variable score level for the polytomous items.

Another way to graphically display the results for the polytomous item (that can also be used for the dichotomous items) is to plot the means of the conditional distributions of item response given matching variable score. Figure 9 gives a plot of the observed conditional item score means and fitted conditional item score means using the polynomial loglinear model of uniform DIF for item 23. The scores on item 23 range from 0 to 3. The lines in Figure 9 give the mean item score as a function of matching variable score. The conditional means are presented separately for males and females. If there were no DIF the conditional means for males and females would be identical. Figure 10 gives the observed and fitted conditional means using the polynomial loglinear model of nonuniform DIF for item 23. The observed conditional means for females are generally below those for males in the middle of the matching score range. The model for uniform DIF (Figure 9) appears to fit the data well. This is consistent with the results in Table 3 which indicated significant uniform DIF, but not significant nonuniform DIF for item 23.

Observed and fitted conditional means using the polynomial loglinear model for uniform DIF are presented in Figure 11 for item 26. The plot of observed and fitted conditional means using the polynomial loglinear model for nonuniform DIF is presented in Figure 12. The model for nonuniform DIF appears to provide a better fit to the data than the model of uniform DIF. This is consistent with the results in Table 3 which indicated significant nonuniform DIF. For matching variable scores above around 19 the means for males are higher than the means for females, whereas for matching variable scores below 19 the opposite is the case. The difference in means between males and females is larger for scores above 19.

Example 2

The second example consists of applying the polynomial log-linear model to investigate DIF for common items in a common-item nonequivalent groups equating design. In the common-item nonequivalent groups equating design the forms of a test to be equated are administered to different groups along with a common set of items. The common items may be included in the score reported to the examinee (an internal set of common items) or not included in the score reported to the examinee (an external set of common items).

For the common item equating to provide valid results it is important that the common items function the same in both test forms. A common item could function differently on two forms due to the different contexts in which it was embedded, or the different times it was administered (the topic of the item might be more salient at one time versus another). One definition of the items functioning the same for both forms is that there is no association between item response and form on which the item was administered when conditioned on the score for all common items. DIF analysis can be used to assess whether this association exists or not. Instead of the focal and reference groups being majority and minority or male and female as is (typical in DIF studies), the groups here are the two forms in which the common item set is embedded and the two test dates on which these two forms were administered.

The data used were from a 150 item professional certification test. The focus was on the 1993 form (administered in 1993). The 1993 form had a link to the 1992 form (administered in 1992 with 37 internal common items to the 1993 form) and the 1991 form (administered in 1991 with 38 internal common items to the 1993 form). There were 1521 examinees who took the 1991 form, 1450 examinees who took the 1992 form, and 1375 examinees who took the 1993 form.

For the 1993/1991 data, d_1 was set equal to four for each studied item after roughly examining how many parameters would be needed to model each item. For each item the likelihood ratio chi-square for testing the nonuniform model versus the saturated model (goodness of fit test) was not significant at the .05 level of significance.

Three different significance levels were used for the analysis — 0.05, 0.01, and 0.05 / 38 = 0.0013158 (a Bonferroni adjustment). For uniform DIF, a total of thirteen items were found to be significant at the .05 level and beyond, ten were significant at the .01 level and beyond, and five were significant at the 0.0013 level of significance. The results for all items are presented in Table 5.

For non-uniform DIF, only two items were significant at the 0.01 level. Both items that exhibited non-uniform DIF also exhibited uniform DIF at the 0.0013 level of significance.

All items that showed significant DIF at the 0.01 level of significance were examined by looking at the actual content of the items (the item stems) and their responses (alternatives). Two of the items for which uniform DIF was indicated at the 0.0013 level of significance had syntactic differences between the two forms. For one of the items, "... NOT..." (all capital letters) was used in the stem while for that item on the other form "...<u>not</u>..." (underlined lower case) was used. For the other item, the word "vs." in the stem was written with a period at the end of the abbreviation on one form, and on the other form it was just written as "vs" without a period at the end. No other noticeable syntactic differences were found for the other items which had significance levels less 0.01. There should be no syntactic differences in an item between forms (every common item should be absolutely identical between forms). These differences were missed by test development staff who checked for the items being identical on the two form.

Items that are functioning differently on the two test forms may have an adverse effect on the equating. To study the effect of the inclusion of the two items with syntactic differences on equating the equating analysis was re-done excluding those two common items. Before proceeding, it was necessary to determine if these two syntactically incorrect items need to stay in the common-item pool for the sake of content specifications. The common-item pool should be a mini-version of the test, and must be balanced in its range of content as similarly as possible to the entire set of items used to compute the score reported to examinees. In this particular test all items fell into one of four content areas. Two of the content areas had large numbers of items; the other two content areas had small numbers of items. The items which exhibited the large amount of uniform DIF, and had syntactic problems, came from content areas in which they could be removed, and no harm would be done to the balance of content in the common-item set.

The equating was re-done excluding these two items as common items. In the recomputed equating the two items are considered non-common items. Tucker and Levine Observed Score equating functions were computed (Kolen and Brennan, 1987). As an indication of the difference in the equatings the number of examinees whose scale scores would change if the two items were not used as common items was calculated (the scale scores range from 0 to 150). For the Tucker equating 23 out of the 1375 examinees who took the new form would have a score change of 1 point (either increase or decrease), and for the Levine equating scores for 310 examinees would have changed by one point. Given that the maximum score change is one point on a 151 point scale and the number of examinees with a one point change is not large, it is concluded that not including these two items as common items does not have an important effect on the equating results.

For the 1993/1992 data, d_1 was also set equal to four for each studied item. For each item the likelihood ratio chi-square for testing the nonuniform model versus the saturated model (goodness of fit test) was not significant at the .05 level of significance.

Again, three different significance levels were used for the analysis — 0.05, 0.01, and 0.05 / 37 = 0.0013514 (a Bonferroni adjustment). For uniform DIF, a total of eight items were found to be significant at the 0.05 level and beyond, four were significant at the 0.01 level and beyond, and none were significant at the 0.0014 level of significance. For non-uniform DIF, only two items were significant at the 0.01 level. Neither of these items exhibited significant nonuniform DIF. The results for the 1993/1992 equating are presented in Table 6.

As in the 1993/1991 equating study, all items that showed significant DIF at the .01 level of significance (and beyond) were examined by looking at the actual content of the items (the item stems) and their responses (alternatives). None of the items manifested any apparent reasons why they should perform differently in the two different forms.

The results indicated more DIF for the 1993/1991 equating items than for the 1993/1992 equating items. A plausible ad-hoc explanation is that since some of the items had somewhat political and time related content, there would be less bias when the time interval between administrations was smaller as any effect of time related content would be reduced.

Discussion

The focus of the investigation of DIF is the conditional association between item response and group given a matching variable. This association can be modeled by loglinear models, or logit models using either the item response or group as the dependent variable.

Loglinear and logit models for studying DIF were presented and loglinear formulations of the logit models were given. A polynomial loglinear model using scores on the matching variable and item responses was introduced. This model contains far fewer parameters than loglinear models that treat the matching variable and item response as nominal. Unlike the logit models, the polynomial loglinear model is generalized to the case of more than two item responses and more than two groups. An advantage of the polynomial loglinear model is that it provides a non-saturated model of nonuniform DIF that is able to detect more complex forms of DIF than logit models that have been suggested (Equation 22 versus Equations 11 and 17), although it is possible that the logit models could be expanded to model more complex forms of DIF.

An example of using the polynomial loglinear model to study DIF was given using data from Miller and Spray (1993). The results of the polynomial loglinear model were compared to the LDFA method given by Miller and Spray (1993). The methods were fairly consistent in their identification of uniform DIF. The LDFA model indicated more nonuniform DIF in the items than the polynomial loglinear model. Examination of graphs of the conditional log-odds ratios for item where the LDFA and loglinear models gave different indications of nonuniform DIF indicated that the polynomial log-linear model appeared to fit better than the LDFA model for most items, although the LDFA model did appear to fit better for one of the items.

The results presented for the polynomial loglinear and LDFA models cannot be used to conclude which model is best for the data used, or even if either model is providing accurate results, since the amount of DIF in the items is unknown. The purpose here was to provide an example of the application of the polynomial loglinear model and a comparison of the results to those obtained from the LDFA model. Simulation could be used to study the the absolute and relative performance of the methods.

A second example involved using the polynomial loglinear model to study DIF in common equating items. In this application of DIF techniques items are studied for differential functioning across different forms in which they are embedded and different test dates on which those forms are given. In common item equating it is important that the common items function the same in the forms being equated, and DIF techniques offer a useful set of tools for studying this question. In the example presented, two items for which the test for uniform DIF was significant were found to have syntactic differences between the forms that were undetected by visual examination of the forms by test development staff.

It would be useful to develop confidence bands as in Miller and Spray (1993) for use in graphical displays such as those displayed in the figures. The usefulness of confidence bands is demonstrated in Miller and Spay (1993) where they are used to identify regions of the matching variable for which DIF is present. Confidence bands and significance tests both have the property that smaller amounts of DIF can be detected as significant with larger samples sizes. This can be a problem when the DIF detected as statistically significant is not practically important.

References

Agresti, A. (1990). *Categorical Data Analysis.* New York: John Wiley & Sons.

Haberman, S. J. (1974). Log-linear models for frequency tables with ordered classifications. *Biometrics, 30,* 589-600.

Hanson, B. A. (1991). A comparison of bivariate smoothing methods in common-item equipercentile equating. *Applied Psychological Measurement, 15,* 391-408.

Hanson, B. A. (1992). *Testing for differences in test score distributions using log-linear models.* Paper presented at the annual meeting of the National Council on Measurement in Education (San Francisco, April).

Holland, P. W., & Thayer, D. T. (1987). *Notes on the use of log-linear models for fitting discrete probability distributions* (Research Rep. No. 87-31). Princeton NJ: Educational Testing Service.

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

Kolen, M. J. (1991). Smoothing methods for estimating test score distributions. *Journal of Educational Measurement, 28,* 257-282.

Kolen, M. J., & Brennan, R. L. (1987). Linear equating models for the common-item nonequivalent-populations design. *Applied Psychological Measurement, 11,* 263-277.

Little, R. J. A., & Rubin, D. B. (1994). Test equating from biased samples, with application to the armed services vocational aptitude battery. *Journal of Educational and Behavioral Statistics, 19,* 309-335.

Livingston, S. A. (1993). Small-sample equating with log-linear smoothing. *Journal of Educational Measurement, 30,* 23-39.

Mellenbergh, G. J. (1982). Contingency table models for assessing item bias. *Journal of Educational Statistics, 7,* 105-118.

Meredith, W., & Millsap, R. E. (1992). On the misuse of manifest variables in the detection of measurement bias. *Psychometrika, 57,* 289-311.

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.

Millsap, R. E., & Everson, H. T. (1993). Methodology Review: Statistical approaches for assessing measurement bias. *Applied Psychological Measurement, 17,* 297-334.

Rosenbaum, P. R., & Thayer, D. T. (1987). Smoothing the joint and marginal distributions of scored two-way contingency tables in test equating. *British Journal of Mathematical and Statistical Psychology, 40,* 43-49.

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

Zwick, R. (1990). When do item response function and Mantel-Haenszel definitions of differential item functioning coincide?, *Journal of Educational Statistics, 15*, 185-197.

Table 1

Haberman Procedure for Item 7 and Item 23

	d1	d2	d3	model: chi-square	model: degrees of freedom	model: p-value	comparison of model to previous model: chi-square	comparison: degees of freedom	comparison: p-value	probability levels for choosing a model
item 7										
	6	1	1	150.17660	166	0.80532				
	5	1	1	150.72087	168	0 82648	0.54426	2	0.76175	
	4	1	1	151.74217	170	0.83928	1.02130	2	0.60011	
	3	1	1	162.43502	172	0.68775	10.69285	2	0.00477	$p^* = 1 - (1 - .01)^{1/5} = .002$
	2	1	1	189.32931	174	0.20207	26.89429	2	0.00000	
	1	1	1	2085.85042	176	0.00000	1896.52912	2	0.00000	
item 23										
	6	3	5	299.11140	322	0.81536				$p^* = 1 - (1 - .003)^{1/4} = .00075$
	6	3	4	320.46034	324	0.54512	21.34894	2	0.00002	
	6	3	3	336.53084	326	0.33211	16.07050	2	0.00032	
	6	3	2	338.00627	328	0.33990	1.47543	2	0.47820	
	6	3	1	344.85630	330	0.27569	6.85003	2	0.03255	
	6	3	5	299.11140	322	0.81536				$p^* = 1 - (1 - .003)^{1/2} = .00150$
	6	2	5	1281.08907	324	0.00000	981.97767	2	0.00000	
	6	1	5	1635.38315	326	0.00000	354.29408	2	0.00000	
	6	3	5	299.11140	322	0.81536				
	5	3	5	299.42433	324	0.83263	0.31293	2	0.85516	
	4	3	5	299.80209	326	0.84815	0.37776	2	0.82789	$p^* = 1 - (1 - .003)^{1/5} = .0006$
	3	3	5	339.18808	328	0.32351	39.38599	2	0.00000	
	2	3	5	361.71406	330	0.11071	22.52599	2	0.00001	
	1	3	5	1926.88129	332	0.00000	1565.16722	2	0.00000	

Table 2

Polynomial Loglinear Models Used For Miller/Spray Open-Ended Items Dataset

item	Number of Parameters		
	d1	d2	d3
22	4	3	4
23	4	3	5
24	4	4	1
25	4	4	2
26	4	4	1
27	4	6	4

Table 3

Polynomial Loglinear Model for the Miller/Spray dataset

	Uniform DIF			Non-Uniform DIF		
	d.f.	chi-square	p <	d.f.	chi-square	p <
Multiple-Choice						
1	1	10.769	0.00103 ***	1	0.022	0.88316
2	1	28.181	0.00000 ***	1	1.777	0.18251
3	1	0.120	0.72931	1	0.268	0.60437
4	1	5.917	0.01499 *	1	1.467	0.22588
5	1	31.676	0.00000 ***	1	0.687	0.40735
6	1	0.763	0.38239	1	0.676	0.41106
7	1	11.277	0.00078 ***	1	3.839	0.05007
8	1	44.426	0.00000 ***	1	1.223	0.26879
9	1	26.096	0.00000 ***	1	0.207	0.64883
10	1	1.265	0.26073	1	0.262	0.60852
11	1	0.179	0.67220	1	0.016	0.89998
12	1	17.280	0.00003 ***	1	0.039	0.84399
Gridded						
13	1	0.025	0.87368	1	0.001	0.97814
14	1	0.030	0.86241	1	1.212	0.27094
15	1	1.278	0.25820	1	6.260	0.01235 *
16	1	8.061	0.00452 **	1	0.155	0.69348
17	1	3.824	0.05053	1	1.359	0.24364
18	1	0.018	0.89427	1	2.423	0.11958
19	1	3.256	0.07114	1	1.008	0.31529
20	1	6.558	0.01044 *	1	0.478	0.48936
21	1	0.145	0.70332	1	0.625	0.42913
Open-Ended						
22	3	1.772	0.62106	4	7.298	0.12097
23	3	18.229	0.00039 ***	5	6.947	0.22463
24	4	5.555	0.23491	1	2.208	0.13732
25	4	6.208	0.18412	2	0.830	0.66045
26	4	15.057	0.00458 **	1	6.098	0.01354 *
27	6	13.382	0.03735 *	4	7.318	0.12002

* <= .05, ** <= .01, *** <= (.05 / 27)

Note-- all dichtomous items were fit with $d_1 = 4$, except for items 7 and 13 where $d_1 = 3$

Table 4

Logistic Discriminant Function Analysis for Miller/SprayDataset

	Uniform DIF			Non-Uniform DIF		
	d.f.	Chi-square	p <	d.f.	Chi-square	p <
Multiple-Choice						
1	1	9.598	0.00195 **	1	1.389	0.23857
2	1	24.595	0.00000 ***	1	1.155	0.28240
3	1	0.059	0.80843	1	2.562	0.10948
4	1	5.322	0.02106 *	1	8.151	0.00430 **
5	1	31.648	0.00000 ***	1	3.924	0.04760 *
6	1	0.825	0.36387	1	4.437	0.03517 *
7	1	13.314	0.00026 ***	1	0.216	0.64189
8	1	43.711	0.00000 ***	1	8.276	0.00402 **
9	1	27.397	0.00000 ***	1	1.289	0.25616
10	1	1.122	0.28955	1	1.246	0.26428
11	1	0.143	0.70578	1	0.112	0.73795
12	1	19.614	0.00001 ***	1	0.256	0.61315
Gridded						
13	1	0.094	0.75918	1	3.141	0.07634
14	1	0.009	0.92439	1	1.193	0.27479
15	1	1.202	0.27299	1	0.008	0.92743
16	1	6.471	0.01097 *	1	4.694	0.03028 *
17	1	3.809	0.05098	1	0.332	0.56465
18	1	0.012	0.91453	1	0.032	0.85901
19	1	5.216	0.02238 *	1	0.003	0.95462
20	1	3.489	0.06177	1	4.626	0.03150 *
21	1	0.076	0.78290	1	0.126	0.72258
Open-Ended						
22	1	1.683	0.19457	1	0.174	0.67647
23	1	17.086	0.00004 ***	1	3.662	0.05566
24	1	3.449	0.06329	1	8.189	0.00421 **
25	1	1.468	0.22572	1	5.956	0.01467 *
26	1	6.813	0.00905 **	1	14.169	0.00017 ***
27	1	5.214	0.02241 *	1	12.318	0.00045 ***

* <= .05, ** <= .01, *** <= (.05 / 27)

Table 5

Polynomial Loglinear Model for the Equating 1993/1991 Dataset

	Uniform DIF			Non-Uniform DIF		
item	d.f.	chi-square	p <	d.f.	chi-square	p <
1	1	0.351	0.55333	1	0.902	0.34212
2	1	2.340	0.12610	1	0.003	0.95570
3	1	3.725	0.05360	1	1.267	0.26031
4	1	1.005	0.31609	1	0.200	0.65494
5	1	7.656	0.00566 **	1	0.046	0.83047
6	1	15.761	0.00007 ***	1	0.062	0.80312
7	1	5.965	0.01460 *	1	0.000	0.99691
8	1	0.881	0.34798	1	2.239	0.13456
9	1	0.646	0.42140	1	0.927	0.33572
10	1	3.019	0.08228	1	1.454	0.22787
11	1	0.145	0.70345	1	0.231	0.63048
12	1	9.378	0.00220 **	1	2.930	0.08697
13	1	0.059	0.80771	1	1.005	0.31601
14	1	15.918	0.00007 ***	1	0.053	0.81859
15	1	5.355	0.02067 *	1	3.426	0.06417
16	1	3.534	0.06010	1	1.342	0.24663
17	1	1.634	0.20116	1	1.003	0.31648
18	1	5.064	0.02443 *	1	1.126	0.28857
19	1	0.620	0.43117	1	0.031	0.86057
20	1	0.176	0.67469	1	0.208	0.64831
21	1	0.219	0.63999	1	0.114	0.73533
22	1	9.056	0.00262 **	1	1.060	0.30321
23	1	0.000	0.98877	1	0.380	0.53749
24	1	2.610	0.10616	1	2.513	0.11291
25	1	7.707	0.00550 **	1	1.465	0.22613
26	1	7.672	0.00561 **	1	0.292	0.58902
27	1	1.890	0.16918	1	1.196	0.27403
28	1	1.265	0.26070	1	0.007	0.93382
29	1	95.091	0.00000 ***	1	8.596	0.00337 **
30	1	24.799	0.00000 ***	1	7.184	0.00735 **
31	1	3.255	0.07122	1	0.972	0.32408
32	1	0.291	0.58980	1	0.219	0.63959
33	1	15.929	0.00007 ***	1	0.549	0.45860
34	1	2.788	0.09497	1	1.178	0.27771
35	1	0.784	0.37600	1	1.695	0.19298
36	1	0.002	0.96228	1	0.013	0.91011
37	1	1.613	0.20402	1	2.020	0.15525
38	1	1.493	0.22175	1	0.200	0.65452

* <= .05, ** <= .01, *** <= (.05 / 38)

Table 6

Polynomial Loglinear Model for the Equating 1993/1992 Dataset

	Uniform DIF			Non-Uniform DIF		
item	d.f.	Chi-square	p <	d.f.	Chi-square	p <
1	1	5.209	0.02248 *	1	0.664	0.41532
2	1	7.616	0.00578 **	1	2.941	0.08634
3	1	1.198	0.27376	1	0.012	0.91144
4	1	1.801	0.17962	1	2.738	0.09798
5	1	0.280	0.59663	1	0.172	0.67793
6	1	0.237	0.62672	1	0.007	0.93153
7	1	0.349	0.55443	1	0.908	0.34053
8	1	0.828	0.36274	1	0.108	0.74214
9	1	9.076	0.00259 **	1	3.238	0.07196
10	1	0.410	0.52216	1	0.565	0.45223
11	1	2.688	0.10114	1	0.000	0.99542
12	1	0.181	0.67060	1	0.688	0.40675
13	1	0.097	0.75553	1	0.578	0.44710
14	1	3.161	0.07543	1	1.346	0.24600
15	1	0.452	0.50118	1	3.973	0.04623 *
16	1	0.341	0.55918	1	0.635	0.42564
17	1	6.886	0.00869 **	1	1.602	0.20563
18	1	3.904	0.04816 *	1	3.015	0.08251
19	1	1.476	0.22433	1	0.111	0.73919
20	1	0.605	0.43657	1	0.807	0.36892
21	1	1.680	0.19494	1	0.110	0.74022
22	1	0.738	0.39025	1	1.284	0.25708
23	1	0.971	0.32445	1	1.460	0.22689
24	1	0.231	0.63087	1	1.466	0.22594
25	1	0.954	0.32868	1	0.269	0.60374
26	1	5.595	0.01801 *	1	0.463	0.49606
27	1	0.450	0.50239	1	3.515	0.06081
28	1	2.687	0.10119	1	4.696	0.03023 *
29	1	6.081	0.01367 *	1	0.028	0.86797
30	1	6.794	0.00915 **	1	0.250	0.61731
31	1	0.229	0.63260	1	1.012	0.31444
32	1	0.017	0.89563	1	0.358	0.54954
33	1	1.099	0.29447	1	0.013	0.90782
34	1	1.753	0.18544	1	0.003	0.95455
35	1	3.164	0.07528	1	0.148	0.70059
36	1	3.621	0.05705	1	0.211	0.64597
37	1	0.224	0.63598	1	0.767	0.38105

* <= .05, ** <= .01, *** <= (.05 / 37)

Figure 1

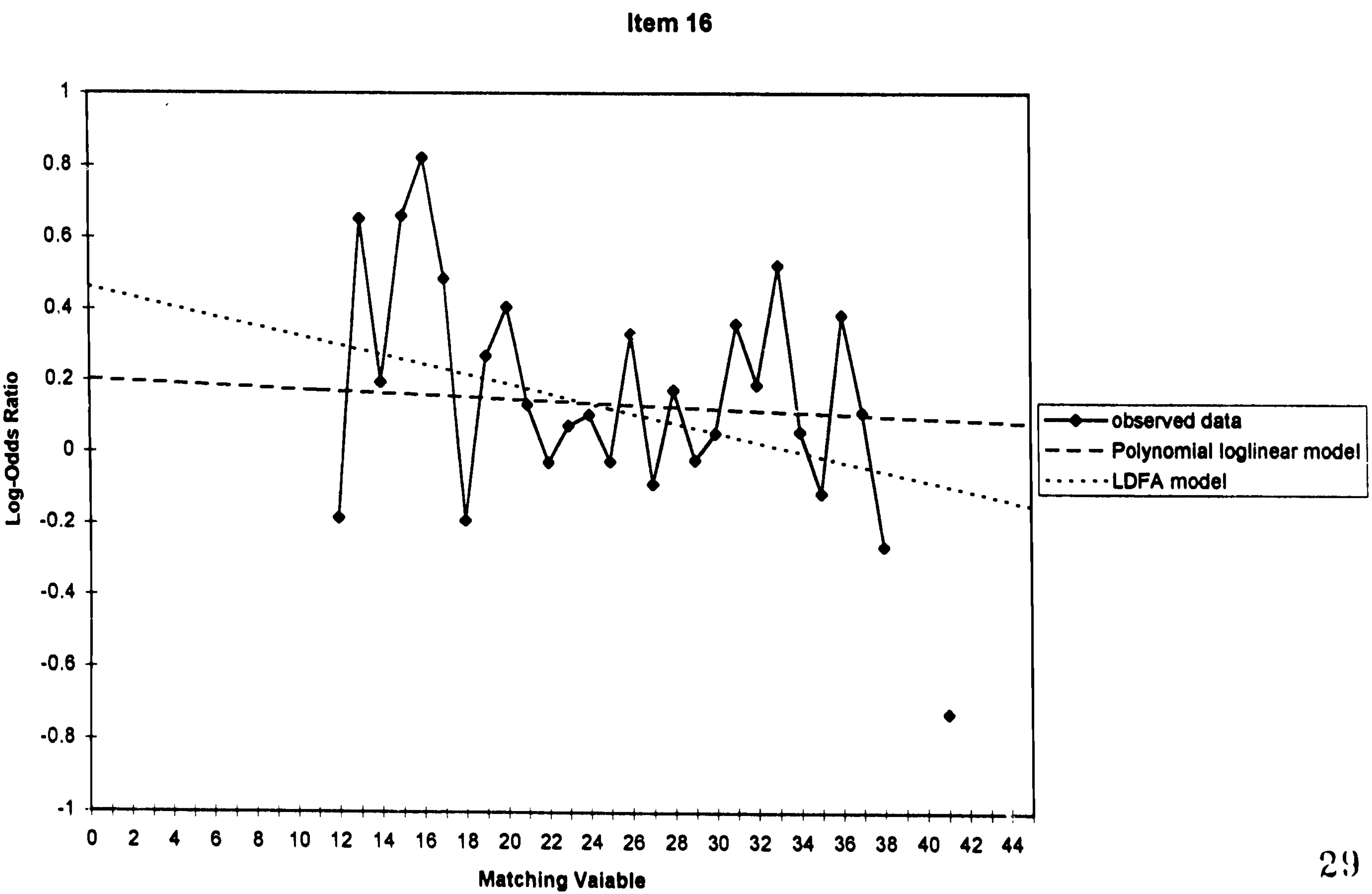

Item 16
Log-Odds Ratio
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1
0 2 4 6 8 10 12 14 16 18 20 22 24 26 28 30 32 34 36 38 40 42 44
Matching Valable
observed data
Polynomial loglinear model
LDFA model

Figure 2

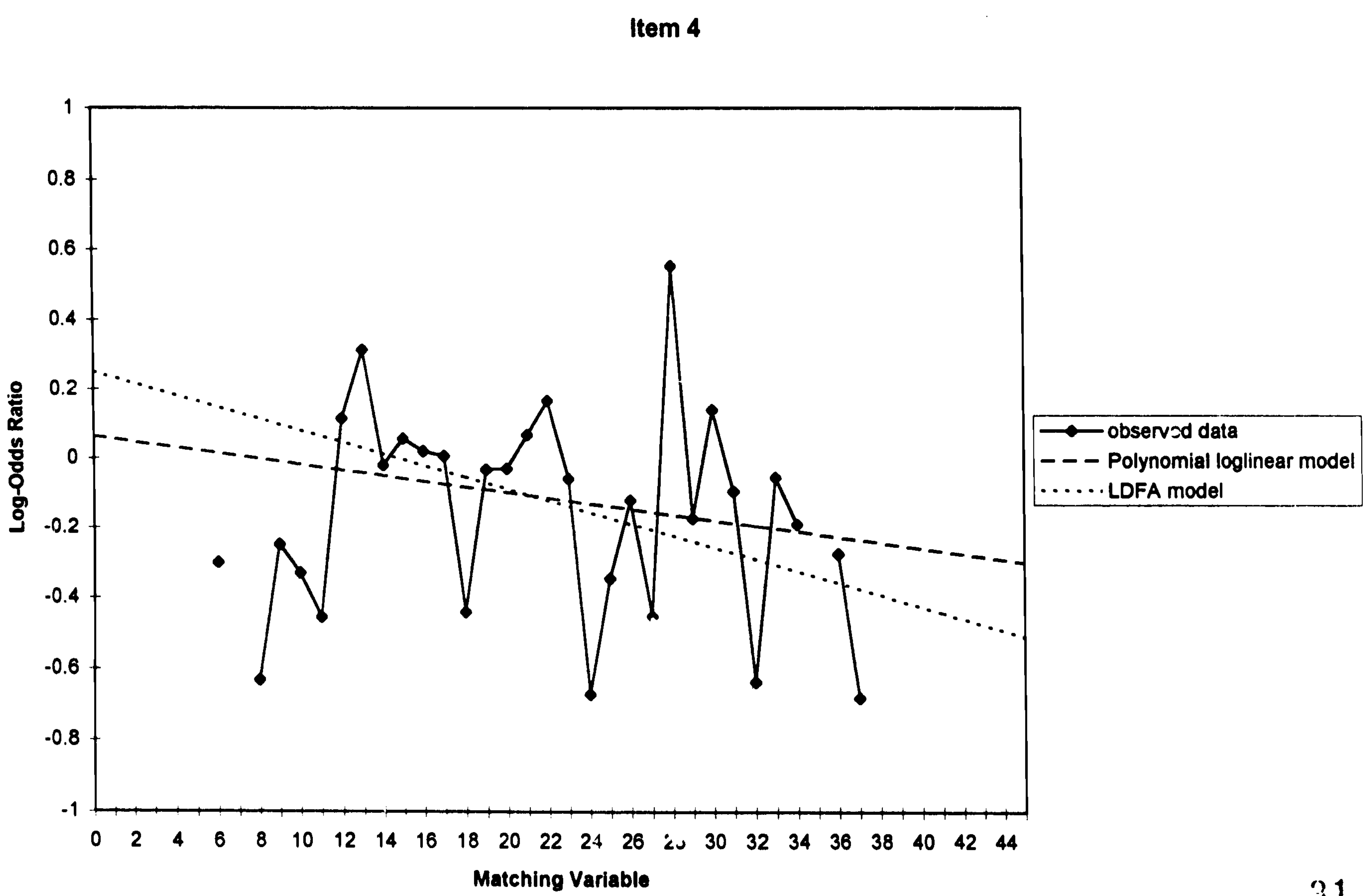
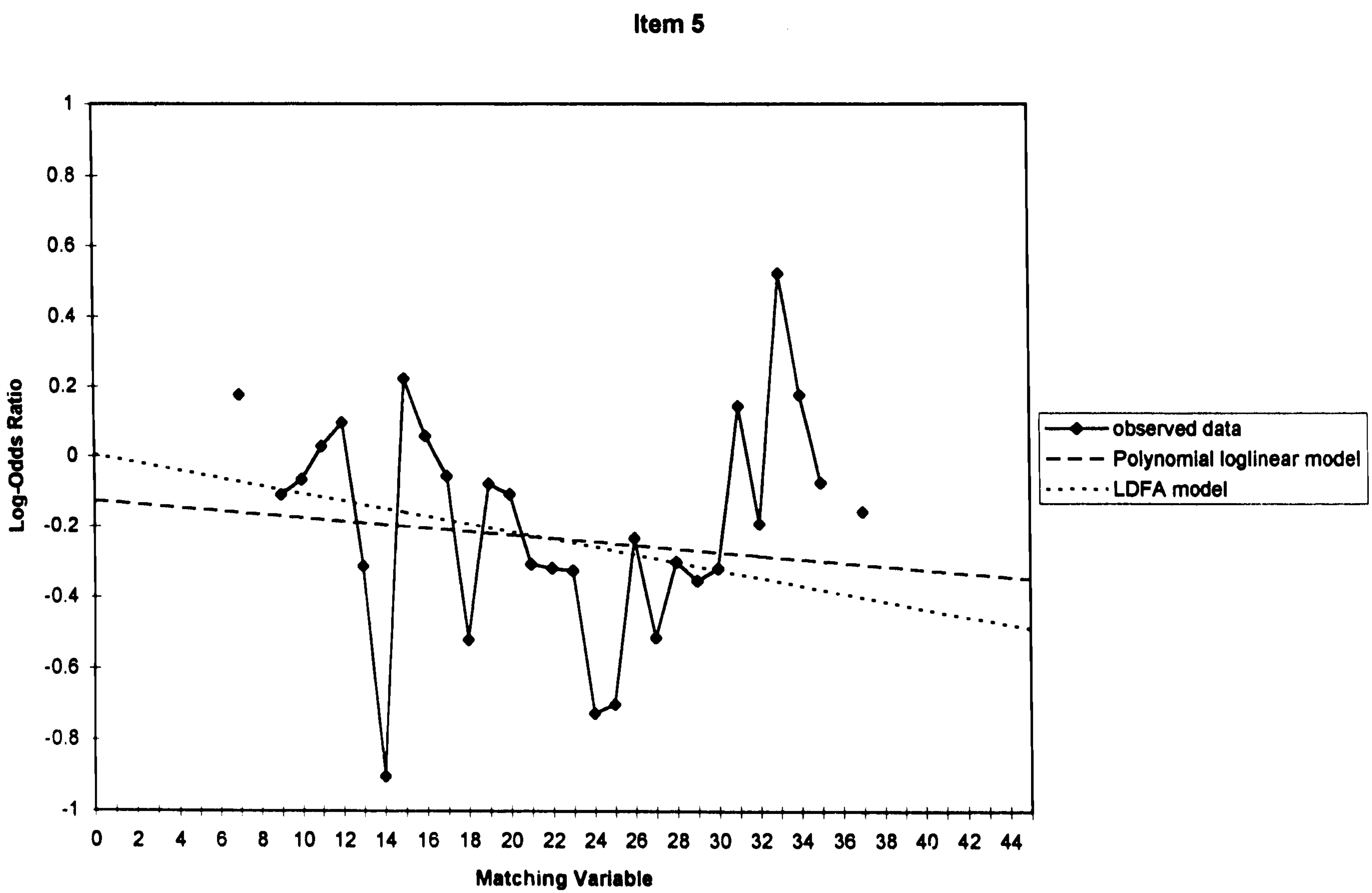

Item 4
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1
Log-Odds Ratio
0 2 4 6 8 10 12 14 16 18 20 22 24 26 28 30 32 34 36 38 40 42 44
Matching Variable
observed data
Polynomial loglinear model
LDFA model

Figure 3


Item 5
Log-Odds Ratio
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1
0
2
4
6
8
10
12
14
16
18
20
22
24
26
28
30
32
34
36
38
40
42
44
Matching Variable
observed data
Polynomial loglinear model
LDFA model

32

Figure 4

Item 6

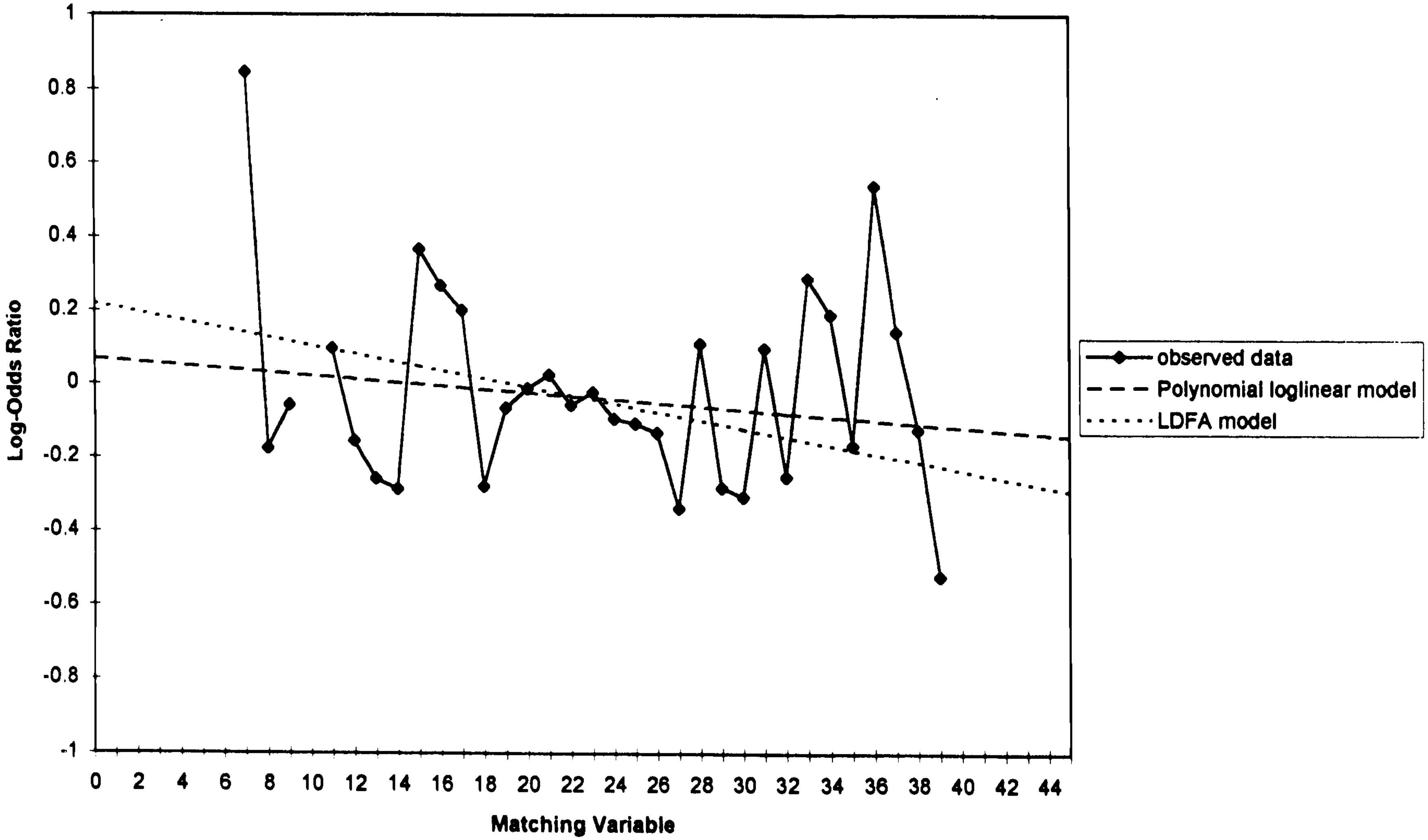


Figure 5

Item 8

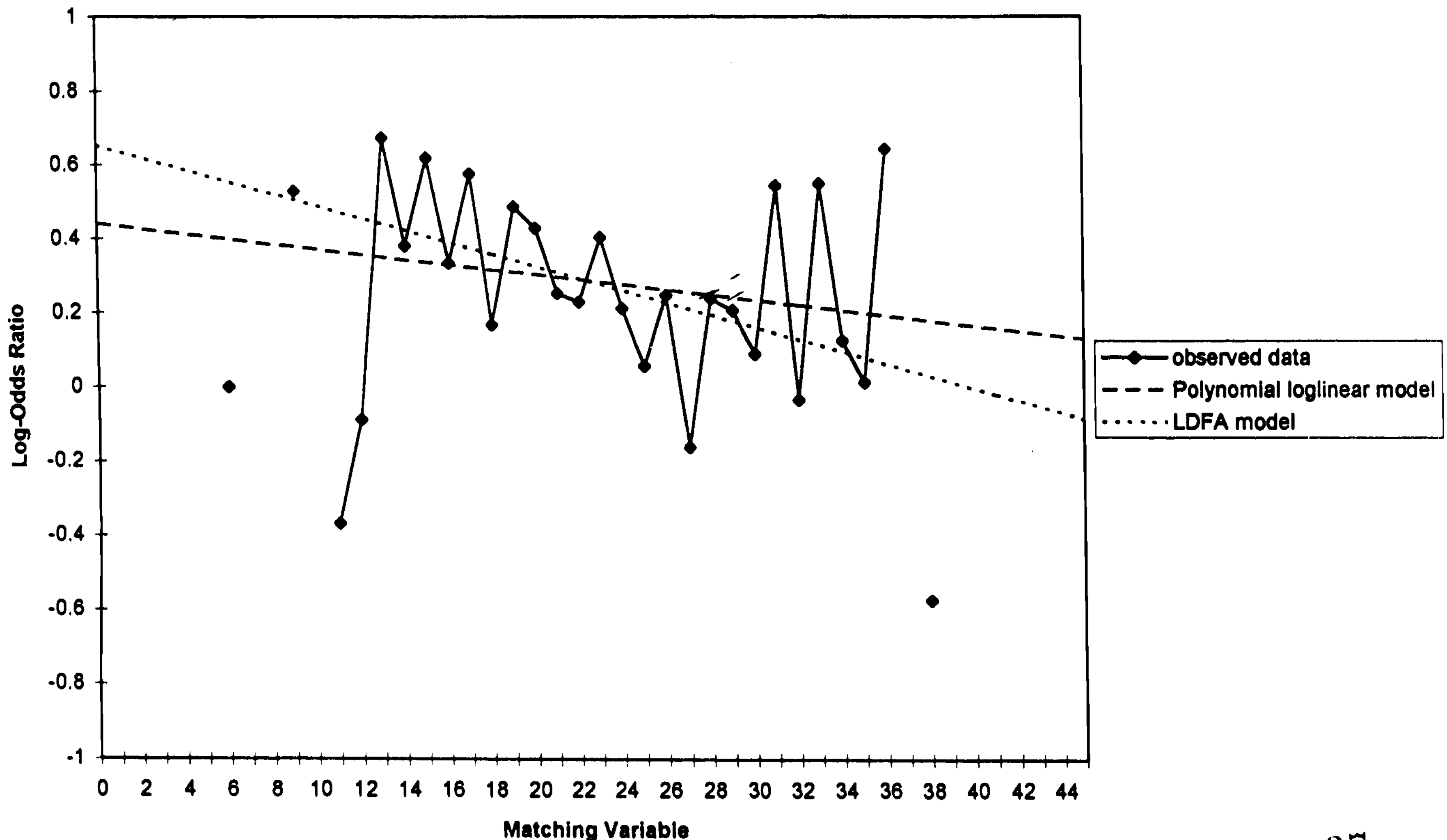


Figure 6

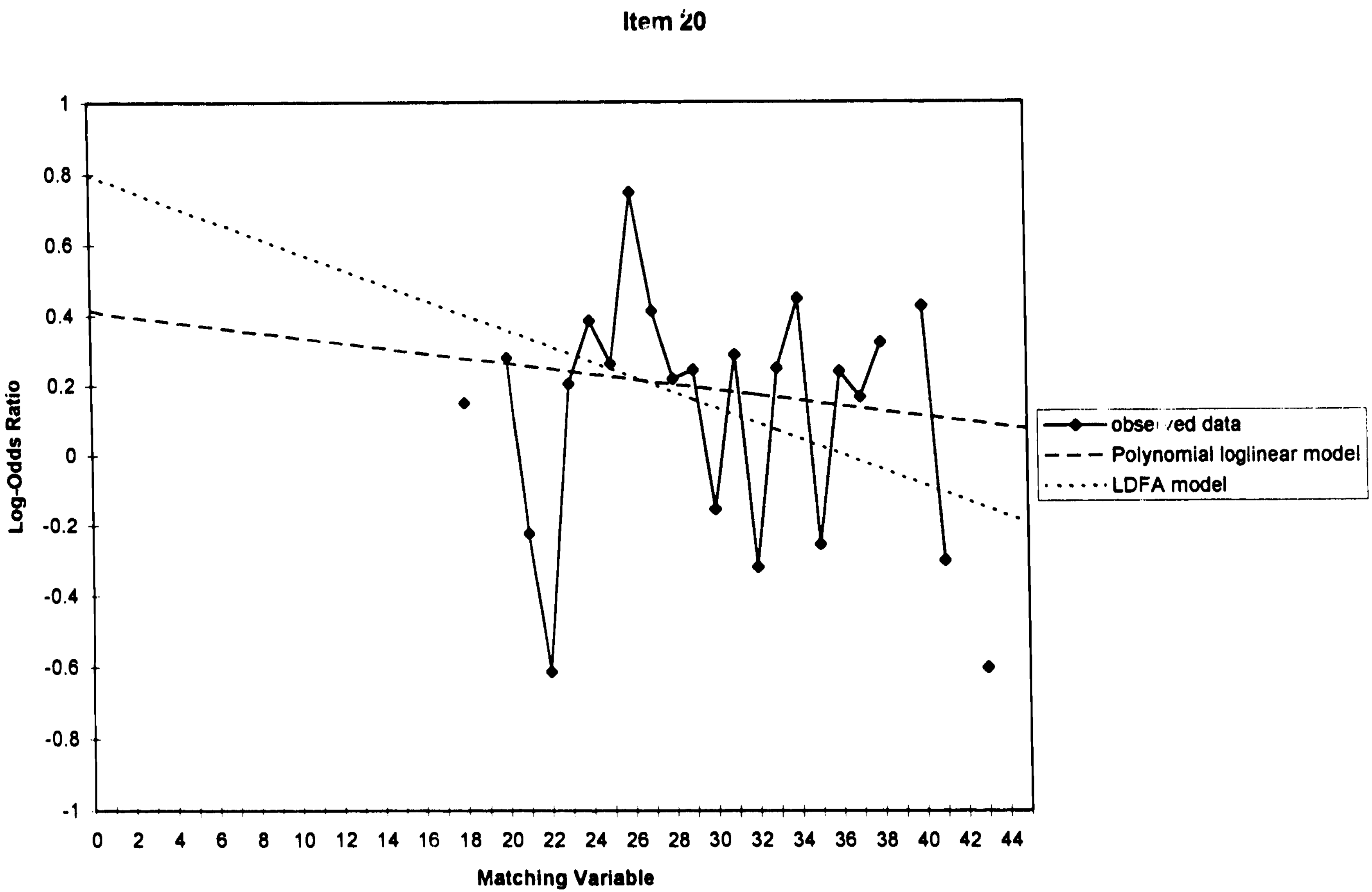

Item 20
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1
Log-Odds Ratio
0 2 4 6 8 10 12 14 16 18 20 22 24 26 28 30 32 34 36 38 40 42 44
Matching Variable
obse ved data
Polynomial loglinear model
LDFA model

Figure 7

Item 15

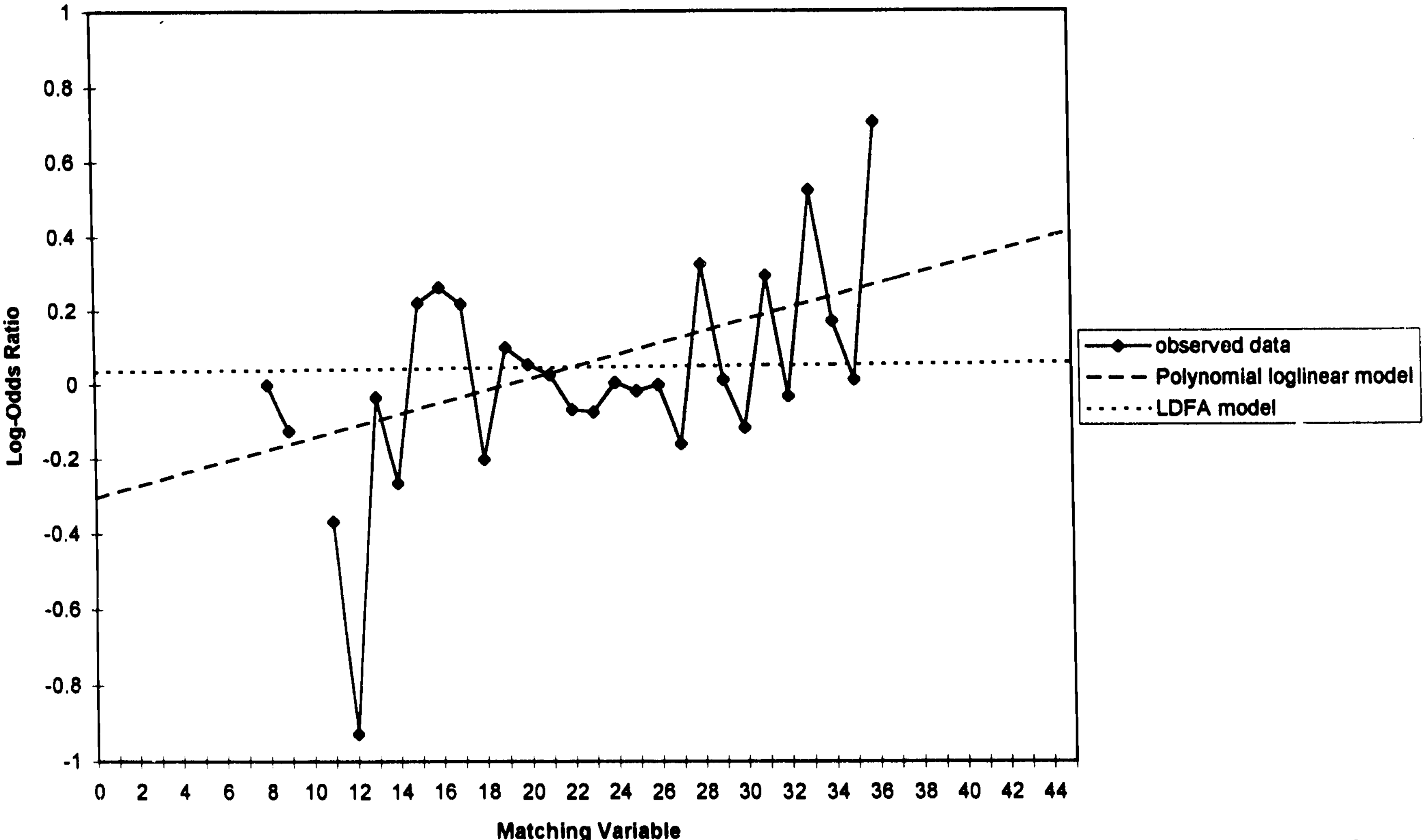


Figure 8

Item 7

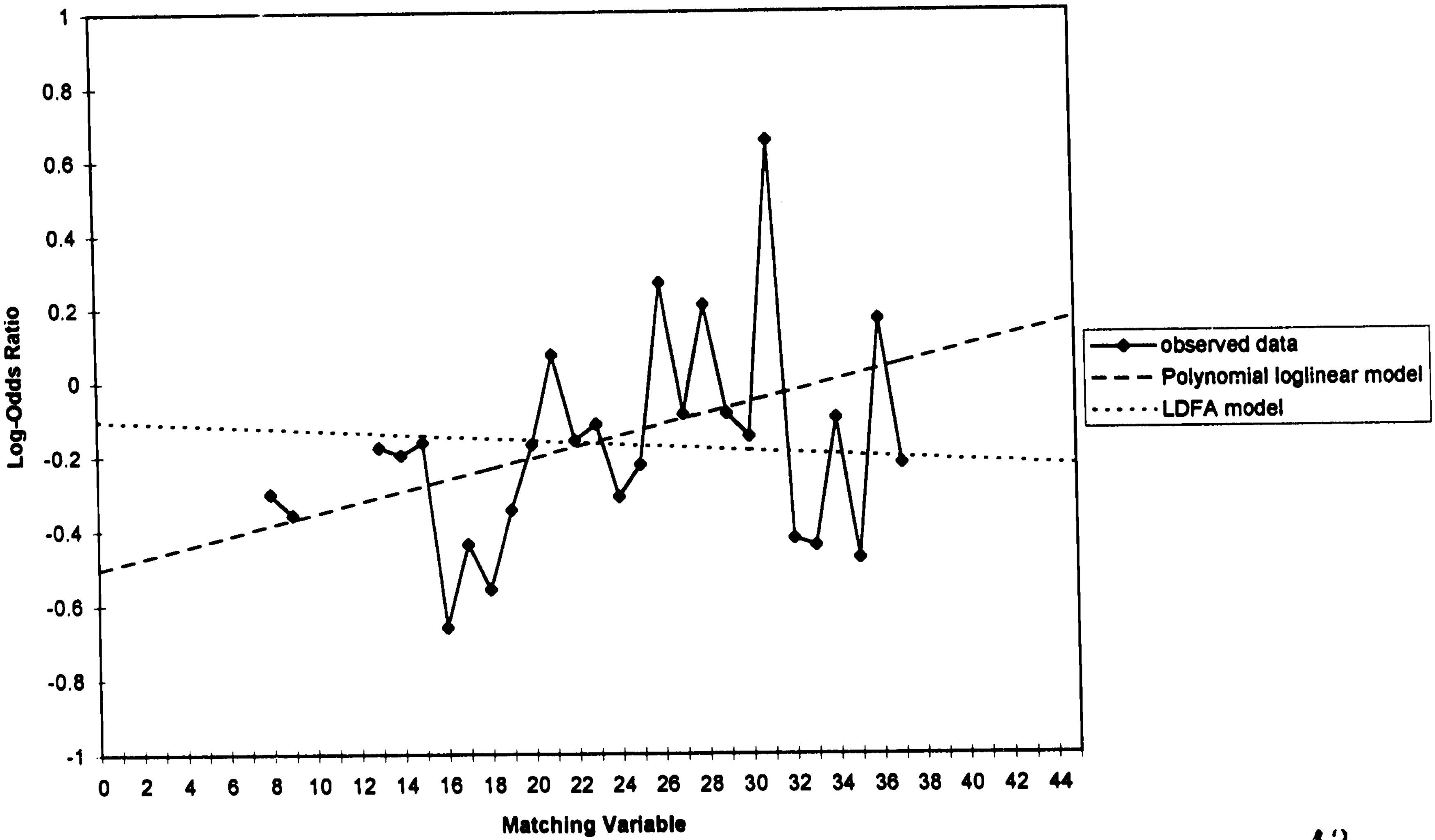
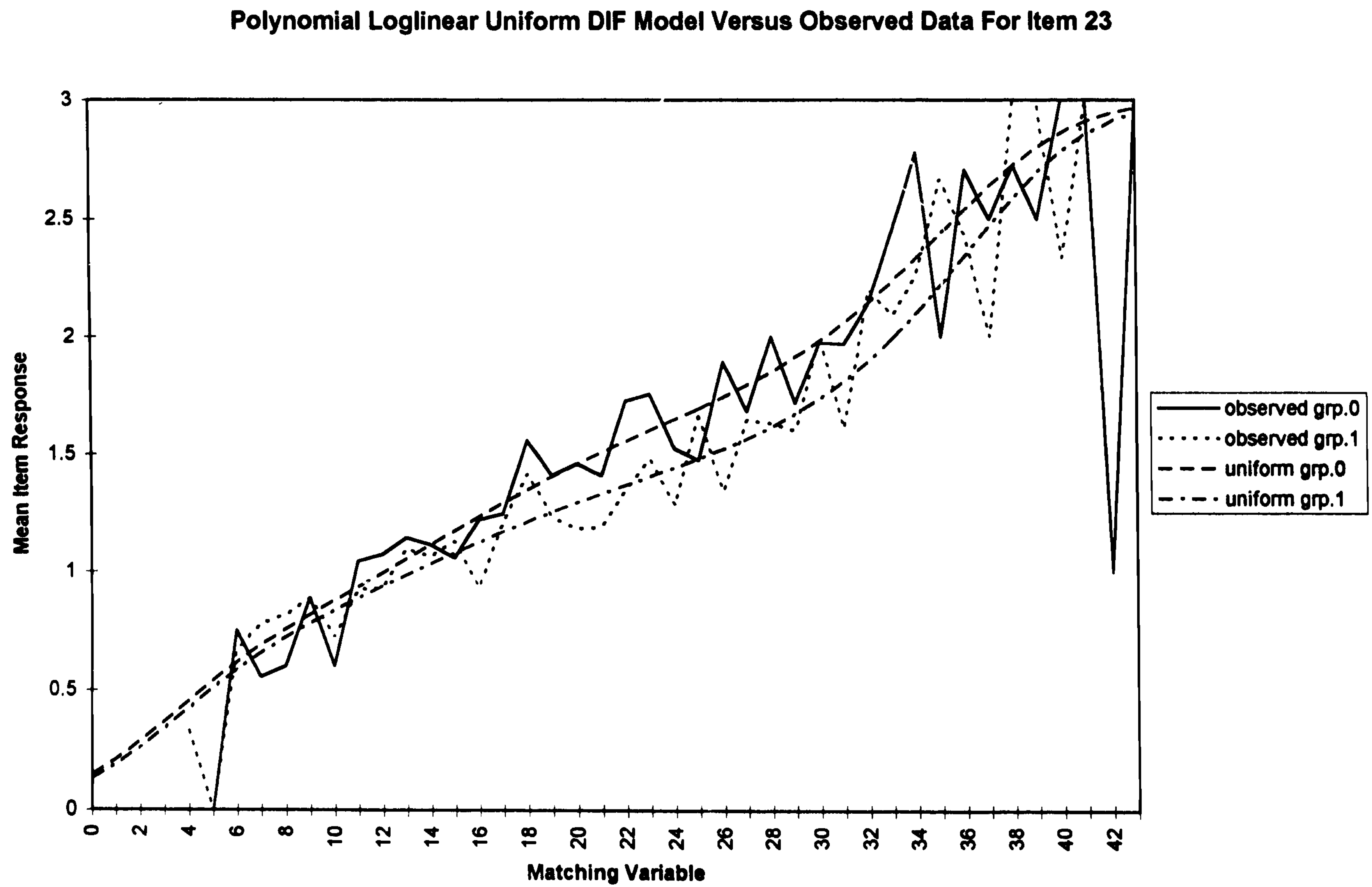


Figure 9



Polynomial Loglinear Uniform DIF Model Versus Observed Data For Item 23
Mean Item Response
3
2.5
2
1.5
1
0.5
0
0
2
4
6
8
10
12
14
16
18
20
22
24
26
28
30
32
34
36
38
40
42
Matching Variable
observed grp.0
observed grp.1
uniform grp.0
uniform grp.1

Figure 10

Polynomial Loglinear Non-Uniform DIF Model Versus Observed Data For Item 23

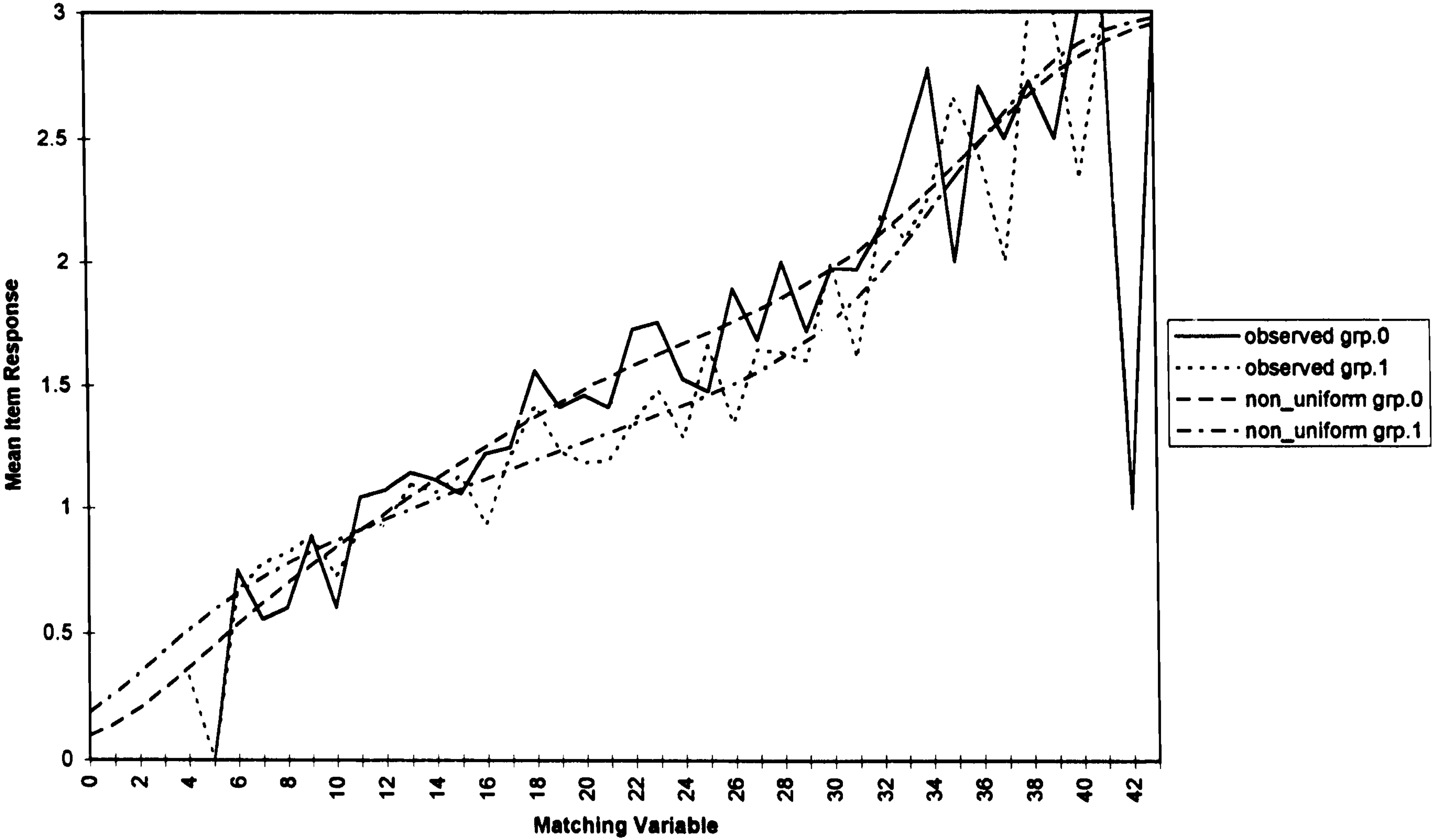


Figure 11

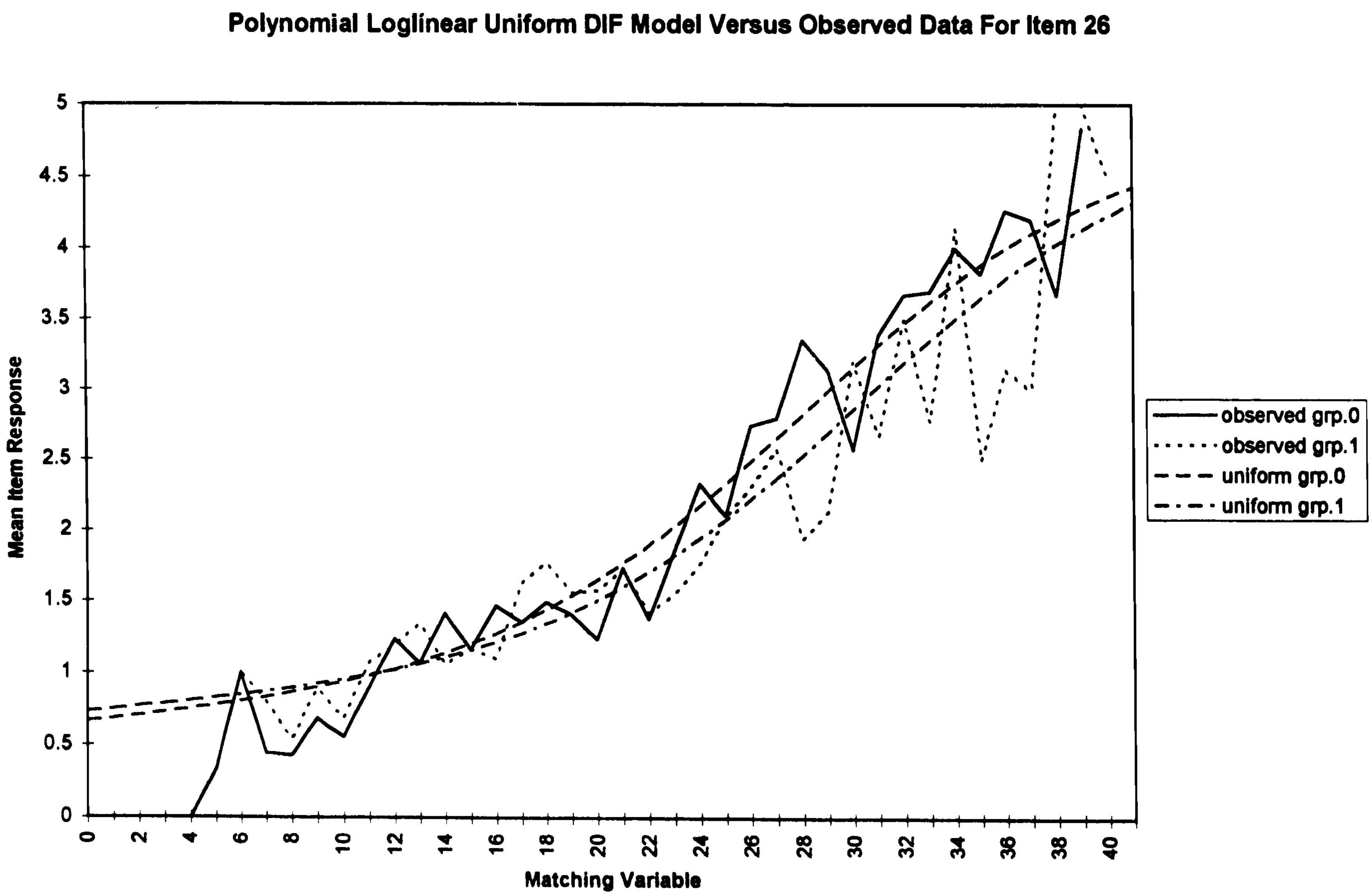

Polynomial Loglinear Uniform DIF Model Versus Observed Data For Item 26
Mean Item Response
0
0.5
1
1.5
2
2.5
3
3.5
4
4.5
5
Matching Variable
0
2
4
6
8
10
12
14
16
18
20
22
24
26
28
30
32
34
36
38
40
observed grp.0
observed grp.1
uniform grp.0
uniform grp.1

Figure 12

Polynomial Loglinear Non-Uniform DIF Model Versus Observed Data For Item 26