ACT Research Report Series 97-1

Application of a Polynomial Loglinear Model to Assessing Differential Item Functioning for Common Items in the Common-Item Equating Design

Bradley A. Hanson

Zachary S. Feinstein

ACT

March 1997

For additional copies write:
ACT Research Report Series
PO Box 168
Iowa City, Iowa 52243-0168

Application of a Polynomial Loglinear Model to Assessing Differential Item Functioning for Common Items in the Common-Item Equating Design

Bradley A. Hanson
Zachary S. Feinstein

Errata for

Application of a Polynomial Loglinear Model to Assessing Differential Item Functioning for Common Items in the Common-Item Equating Design

ACT Research Report Series 97-1

Bradley A. Hanson
Zachary S. Feinstein

The third line after Equation 6 on page 3 should read as follows:

at each level of the matching variable. If $\theta_{11k} = 1$ then group and item response are independent

Abstract

Loglinear and logit models that have been suggested for studying differential item functioning (DIF) are reviewed, and loglinear formulations of the logit models are given. A polynomial loglinear model for assessing DIF is introduced which incorporates scores on the matching variable and item responses. The polynomial loglinear model contains far fewer parameters than loglinear models which treat the matching variable and item response as nominal. Compared to logit models that have been presented for investigating DIF, the polynomial loglinear model is easier to generalize to the case of more than two groups and more the two response categories, and can model more complex forms of DIF. The use of DIF methodology to investigating whether common items in the common-item equating design are functioning differently across test forms is discussed. Examples are given of using the polynomial loglinear model to investigate DIF for a test containing dichotomous and polytomous items, and for investigating DIF for common items in a common-item nonequivalent groups equating design.

Acknowledgement. The authors thank Mary Pommerich and Ron Cope for helpful comments on an earlier version of this paper, and thank Judy Spray for providing the data used in one of the examples.

Application of a Polynomial Loglinear Model to Assessing Differential Item Functioning for Common Items in the Common-Item 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 examinees the test is intended for. Procedures of this type are part of the process of construct validation. One aspect of investigating the validity of a test for various groups of examinees is the investigation of the test items for differential item functioning. Differential item functioning (DIF) is said to exist when an item is functioning differently for two or more groups of examinees, within the population the test is intended for. DIF manifests itself by differential response to an item based on the group an examinee belongs to, when conditioned on the latent variable being measured by the test the item is a part of. Differential item functioning is defined conditioned on the latent variable measured by the test. This is in contrast to differences in responding to an item among groups when averaged across levels of the latent variable. These marginal differences in item performance may reflect legitimate differences between the groups on the latent variable measured by the test (denoted impact), and do not necessarily represent DIF.

The first part of this paper presents a definition of DIF and reviews loglinear and logit models used for assessing DIF. Loglinear formulations of logit models that have been suggested for studying DIF are presented.

The second part of this paper presents a polynomial loglinear model for assessing DIF which incorporates numerical scores for the item response variable and conditioning variable. An example using the polynomial loglinear model for investigating DIF on a test with both dichotomous and polytomous items is given.

The third part of this paper discusses applying DIF methodology to investigating whether common items in a common-item equating design are functioning differently across test forms. In the common-item equating design the forms of a test to be equated are administered to different groups along with a common set of items. For common-item equating to provide valid results it is important that the common items function the same on all test forms. DIF techniques can be used to investigate whether common items are functioning differently on different forms of a test. An example is presented applying the polynomial loglinear model discussed in the second part of the paper to investigate DIF for common items in a 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 variable 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 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).

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). DIF defined using an observed matching variable may not correspond to DIF defined using the latent variable measured by the items as the matching variable (the true matching variable). The relationship between DIF defined using the true matching variable and an observed matching variable is discussed by Holland and Thayer (1988), Zwick (1990), and Meredith and Millsap (1992).

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

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

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

As an example of how to interpret an odds ratio consider the case of two groups and two item response categories. Item response category 1 represents a correct response and item response category 2 represents an incorrect response. Let $p_{ij|k} = m_{ijk}/N_k$, where N_k is the number of examinees with a matching variable score in category k, so $p_{ij|k}$ is the probability of observing an examinee in group j with an item response in category i conditioned on the matching variable score being in category k. The odds of having a correct response versus an incorrect response for examinees in group 1 with a matching variable score in category k is the conditional probability of a correct response for examinees in group 1 having a matching variable score in category k divided by the conditional probability of an incorrect response for examinees in group 1 having a matching variable score in category k. This odds ratio is given by

$$\frac{p_{11|k}/(p_{11|k} + p_{21|k})}{p_{21|k}/(p_{11|k} + p_{21|k})} = \frac{p_{11|k}}{p_{21|k}} = \frac{m_{11k}/N_k}{m_{21k}/N_k} = \frac{m_{11k}}{m_{21k}}. \tag{4}$$

Similarly, odds of having a correct response versus an incorrect response for examinees in group 2 with a matching variable score in category k is

$$\frac{p_{12|k}/(p_{12|k} + p_{22|k})}{p_{22|k}/(p_{12|k} + p_{22|k})} = \frac{p_{12|k}}{p_{22|k}} = \frac{m_{12k}/N_k}{m_{22k}/N_k} = \frac{m_{12k}}{m_{22k}}. \tag{5}$$

The value of θ_{11k} in Equation 3 is the ratio of the odds of having a correct versus incorrect response in groups 1 and 2 for matching variable category k (Equation 4 divided by Equation 5):

$$\theta_{11k} = \frac{m_{11k}/m_{21k}}{m_{12k}/m_{22k}} = \frac{m_{11k}m_{22k}}{m_{21k}m_{12k}}. \tag{6}$$

In the case of two groups and two item response categories only the one conditional odds ratio given in Equation 6 is needed to describe the relationship between group and item response at each level of the matching variable. If $\theta_{11k} = 0$ then group and item response are independent for matching variable category k. If $\theta_{11k} > 1$ then the odds of getting a correct versus incorrect response is greater for group 1 than group 2 for matching variable category k, and if $\theta_{11k} < 1$ then the odds of getting a correct versus incorrect response is greater for group 2 than group 1 for matching variable category k. With more than two groups or more than two categories of item response more than one odds ratio is needed to describe the relationship between group and item response (Equation 3).

It is sometimes more convenient to use the log of the odds ratios in Equation 3 given by:

$$\log(\theta_{ijk}) = \log(m_{ijk}) + \log(m_{i+1,j+1,k}) - \log(m_{i+1,j,k}) - \log(m_{ij+1,k}), \qquad (7)$$

for $1 \leq i < I$, $1 \leq j < J$. The values in Equation 7 are called the log-odds ratios. The log-odds ratios are symmetric around zero, so a positive and negative log-odds ratio of the same magnitude indicate the same degree of association in opposite directions. No DIF exists if all the log-odds ratios in Equation 7 are equal to zero.

Loglinear and Logit Models for Studying DIF

A similar procedure is used to test for DIF in each of the loglinear and logit models to be discussed. In each case there are three models fit to the data: 1) a model corresponding to nonuniform DIF, 2) a model corresponding to uniform DIF, and 3) a model corresponding to no DIF. The no DIF model is nested within the uniform DIF model, and the uniform DIF model is nested with the nonuniform DIF model.

The likelihood ratio chi-squared statistics for the nonuniform and uniform DIF models are used to test for nonuniform DIF. Under the null hypothesis that the uniform DIF model holds, the difference in the likelihood ratio chi-squared statistics between the uniform and nonuniform DIF models is asymptotically distributed as a chi-square random variable with degrees of freedom equal to the difference in the number of parameters between the two models (call this difference df_n). For a level of significance p, the null hypothesis that the uniform DIF model holds versus the alternative hypothesis that the nonuniform DIF model holds is rejected if the difference in the likelihood ratio chi-square statistics between the uniform and nonuniform DIF models is greater then the upper p percentage point for the chi-square distribution with df_n degrees of freedom.

To test for uniform DIF the likelihood ratio chi-squared statistics for the uniform and no DIF models are used. Under the null hypothesis that the no DIF model holds, the difference in the likelihood ratio chi-squared statistics between the no DIF and uniform DIF models is asymptotically distributed as a chi-square random variable with degrees of freedom equal to the difference in the number of parameters between the two models (df_u). For a level of significance p, the null hypothesis that the no DIF model holds versus the alternative hypothesis that the uniform DIF model holds is rejected if the difference in the likelihood ratio chi-square statistics between the uniform DIF and

no DIF models is greater then the upper p percentage point for the chi-square distribution with df_u degrees of freedom.

Loglinear Models

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

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

One constraint is placed on each of the parameters λ_i^Y, λ_j^V, λ_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 on the λ_{ijk}^{YVZ}, for example $\lambda_{1jk}^{YVZ} = \lambda_{i1k}^{YVZ} = \lambda_{ij1}^{YVZ} = 0$ for all i, j, k. The total number of free parameters is $1+I-1+J-1+K-1+IJ-(I+J-1)+IK-(I+K-1)+JK-(J+K-1)+IJK-(IJ+IK+JK-I-J-K+1) = IJK$. The number of free model parameters equals the number of cells in the table (this is a saturated model). The model in Equation 8 has no residual degrees of freedom (the model fits any data perfectly).

The log-odds ratios in Equation 7 for the model in Equation 8 are

$$\log\left(\frac{m_{ijk}m_{i+1,j+1,k}}{m_{i+1,j,k}m_{i,j+1,k}}\right) = \lambda_{ij}^{YV} + \lambda_{i+1,j+1}^{YV} - \lambda_{i+1,j}^{YV} - \lambda_{i,j+1}^{YV} + \lambda_{ijk}^{YVZ} + \lambda_{i+1,j+1,k}^{YVZ} - \lambda_{i+1,j,k}^{YVZ} - \lambda_{i,j+1,k}^{YVZ}. \quad (9)$$

The log-odds ratios given in Equation 9 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 8 is nonuniform DIF.

Mellenbergh (1982) identifies two nonsaturated models nested within the model given in Equation 8 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 8:

$$\log(m_{ijk}) = \mu + \lambda_i^Y + \lambda_j^V + \lambda_k^Z + \lambda_{ij}^{YV} + \lambda_{ik}^{YZ} + \lambda_{jk}^{VZ}, \quad (10)$$

with the same constraints on the parameters as were indicated for the model in Equation 8. The log-odds ratios in Equation 7 for the model in Equation 10 are

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

The log-odds in Equation 11 will in general differ from 1 but do not differ across levels of the matching variable. Thus, the model given by Equation 10 implies uniform DIF.

The no DIF model presented by Mellenbergh (1982) is obtained by eliminating λ_{ij}^{YV} from the model in Equation 10:

$$\log(m_{ijk}) = \mu + \lambda_i^Y + \lambda_j^V + \lambda_k^Z + \lambda_{ik}^{YZ} + \lambda_{jk}^{VZ}. \tag{12}$$

The log-odds ratios in Equation 7 for the model in Equation 12 will all be zero. Thus, the model in Equation 12 implies no DIF for the item.

To test for nonuniform DIF the procedure described above is used with the model for nonuniform DIF given by Equation 8 and the model for uniform DIF given by Equation 10. The degrees of freedom for testing the null hypothesis of uniform DIF versus the alternative hypothesis of nonuniform DIF is $(I-1)(J-1)(K-1)$. To test for uniform DIF the procedure described above is used with the model for uniform DIF given by Equation 10 and the model for no DIF given by Equation 12. The degrees of freedom for testing the null hypothesis of no DIF versus the alternative hypothesis of uniform DIF is $(I-1)(J-1)$.

Logit Models

Mellenbergh (1982) notes that for dichotomous items logit models equivalent to the loglinear models in Equations 8, 10 and 12 for the purposes of studying DIF can be used. The response variable for the logit models 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 scores 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

$$\log\left(\frac{m_{1jk}}{m_{2jk}}\right) = \alpha_0 + \lambda_j^V + \alpha_1 s_k + \alpha_{2j} s_k \,, \tag{13}$$

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 13 is equivalent to the following loglinear model (Agresti, 1990, pages 152-153)

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

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 8. 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 14, unlike the logit model in Equation 13, it is possible that the number of item response categories could be greater than 2. The log-odds ratios in Equation 7 for the model in Equation 14 are given by

$$\log\left(\frac{m_{ijk}m_{i+1,j+1,k}}{m_{i+1,j,k}m_{i,j+1,k}}\right) = \lambda_{ij}^{YV} + \lambda_{i+1,j+1}^{YV} - \lambda_{i+1,j}^{YV} - \lambda_{i,j+1}^{YV} + (\gamma_{ij} + \gamma_{i+1,j+1} - \gamma_{i+1,j} - \gamma_{i,j+1}) s_k \,. \tag{15}$$

The log-odds ratios in Equation 15 are linear functions of the matching variable score and therefore represent nonuniform DIF.

Eliminating the γ_{ij} terms from the model in Equation 14 gives

$$\log(m_{ijk}) = \mu + \lambda_i^Y + \lambda_j^V + \lambda_k^Z + \lambda_{ij}^{YV} + \lambda_{jk}^{VZ} + \beta_i s_k \,. \tag{16}$$

The log-odds ratios in Equation 7 for the model in Equation 16 are

$$\log\left(\frac{m_{ijk}m_{i+1,j+1,k}}{m_{i+1,j,k}m_{i,j+1,k}}\right) = \lambda_{ij}^{YV} + \lambda_{i+1,j+1}^{YV} - \lambda_{i+1,j}^{YV} - \lambda_{i,j+1}^{YV} \,. \tag{17}$$

Equation 17 may be different from zero but it is constant for all values of the matching variable score. Consequently, the model in Equation 16 represents uniform DIF. The log-odds ratios in

Equations 17 and 11 have the same form since the only difference between the models in Equations 16 and 10 are the interaction terms involving the item response and matching variable which cancel out when computing the odds ratio. Even though the form of the log-odds in Equations 11 and 17 are the same the estimates of the log-odds in the two equations will differ since they are based on different models.

Eliminating the λ_{ij}^{YV} from the model in Equation 16 gives

$$\log(m_{ijk}) = \mu + \lambda_i^Y + \lambda_j^V + \lambda_k^Z + \lambda_{jk}^{VZ} + \beta_i s_k \,. \tag{18}$$

The log-odds ratios in Equation 7 for the model in Equation 18 are all zero. Consequently, the model in Equation 18 represents no DIF.

The models in Equations 14 and 16 are used to test for nonuniform DIF (the test for nonuniform DIF has $I + J - 1$ degrees of freedom). The models in Equations 16 and 18 are used to test for uniform DIF (the test for uniform DIF has $(I - 1)(J - 1)$ degrees of freedom).

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. This logit model can be written as

$$\log\left(\frac{m_{i1k}}{m_{i2k}}\right) = \alpha_0 + \alpha_1 s_k + \alpha_2 r_i + \alpha_3 s_k r_i \,. \tag{19}$$

The logit model in Equation 19 can be written as the following loglinear model

$$\log(m_{ijk}) = \mu + \lambda_i^Y + \lambda_j^V + \lambda_k^Z + \lambda_{ik}^{YZ} + \beta_{1j} s_k + \beta_{2j} r_i + \gamma_j s_k r_i \,. \tag{20}$$

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 8. One constraint is put on each of the parameters β_{1j}, β_{2j} and γ_j (for example, $\beta_{11} = \beta_{21} = \gamma_1 = 0$). For the model in Equation 20 the log-odds ratios in Equation 7 are

$$\log\left(\frac{m_{ijk}m_{i+1,j+1,k}}{m_{i+1,j,k}m_{i,j+1,k}}\right) = (\beta_{2,j+1} - \beta_{2j})(r_{i+1} - r_i) + (\gamma_{j+1} - \gamma_j)(r_{i+1} - r_i)s_k \,. \tag{21}$$

The log-odds ratios in Equation 21 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 20.

Eliminating the terms involving γ_j from the model in Equation 20 gives

$$\log(m_{ijk}) = \mu + \lambda_i^Y + \lambda_j^V + \lambda_k^Z + \lambda_{ij}^{YV} + \beta_{1j}s_k + \beta_{2j}r_i\,. \tag{22}$$

The log-odds ratios in Equation 7 for the model in Equation 22 are

$$\log\left(\frac{m_{ijk}m_{i+1,j+1,k}}{m_{i+1,j,k}m_{i,j+1,k}}\right) = (\beta_{2,j+1} - \beta_{2j})(r_{i+1} - r_i)\,. \tag{23}$$

The log-odds ratios in Equation 23 may differ from zero, but do not vary with the matching variable score. Consequently, uniform DIF is implied by the model in Equation 22.

Eliminating the β_{2j} from Equation 22 gives

$$\log(m_{ijk}) = \mu + \lambda_i^Y + \lambda_j^V + \lambda_k^Z + \lambda_{ij}^{YV} + \beta_{1j}s_k\,. \tag{24}$$

The log-odds ratios in Equation 7 for the model in Equation 24 will all be zero. Consequently, no DIF is implied by Equation 24.

The models in Equations 20 and 22 are used to test for nonuniform DIF (the test for nonuniform DIF has $J - 1$ degrees of freedom). The models in Equations 22 and 24 are used to test for uniform DIF (the test for uniform DIF has $J - 1$ degrees of freedom). Using the models in Equations 20, 22 and 24 to study DIF is called Logistic Discriminant Function Analysis (LDFA) by Miller and Spray (1993).

An advantage of using the logit form of the models (Equations 13 and 19) as opposed to the loglinear form of these models (Equations 14 and 20) 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 ($I > 2$) and more than 2 groups ($J > 2$) without the complications of having to deal with a polytomous dependent variable. The model in Equation 13 can only be used when there are two item response categories, and the model in Equation 19 can only be used when there are 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 there are in the loglinear forms of the logit models in Equations 14 and 20, or the loglinear models presented in Equations 8, 10 and 12.

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. Examples include 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, 1996). This section presents a model for the three-way table of item response, group, and matching variable that can be used to investigate DIF. The model is analogous to polynomial loglinear models that have been used in other measurement applications.

In the loglinear models in Equations 8, 10, and 12 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

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

where $d_1 < K$, $d_2 < I$. As in Equation 8 a constraint is put on the λ_j^V. There are no constraints placed 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 14 and 20 are not special cases of the model in Equation 25.

The log-odds ratios in Equation 7 for the model in Equation 25 are

$$\log\left(\frac{m_{ijk}m_{i+1,j+1,k}}{m_{i+1,j,k}m_{i,j+1,k}}\right) = \sum_{h=1}^{d_2} (\beta_{2,h,j+1} - \beta_{2,hj})\,(r_{i+1}^h - r_i^h) + \sum_{g=1}^{d_1} \sum_{h=1}^{d_2} (\gamma_{gh,j+1} - \gamma_{ghj})\,(r_{i+1}^h - r_i^h)\, s_k^g \,. \tag{26}$$

Equation 26 represents nonuniform DIF. The DIF given in Equation 26 is constrained relative to the DIF given by the saturated loglinear model (Equation 9). The model in Equation 25 is a nonsaturated loglinear model that allows for nonuniform DIF. Comparing Equation 26 to Equations 15 and 21 it

is seen that the loglinear model in Equation 25 allows for more complicated forms of DIF than the models in Equations 14 and 20.

The constrained version of the model given in Equation 25 which implies uniform DIF is

$$\log(m_{ijk}) = \mu + \lambda_j^V + \sum_{g=1}^{d_1} \beta_{1gj} s_k^g + \sum_{h=1}^{d_2} \beta_{2hj} r_i^h + \sum_{g=1}^{d_1} \sum_{h=1}^{d_2} \gamma_{gh} s_k^g r_i^h . \tag{27}$$

The model in Equation 27 differs from the model in Equation 25 by not having the γ_{gh} parameters differ for the different groups. The difference in the number of parameters between the models in Equations 27 and 25 is $d_3(J-1)$. The log-odds ratios in Equation 7 for the model in Equation 27 are

$$\log\left(\frac{m_{ijk} m_{i+1,j+1,k}}{m_{i+1,j,k} m_{i,j+1,k}}\right) = \sum_{h=1}^{d_2} (\beta_{2,h,j+1} - \beta_{2hj})\,(r_{i+1}^h - r_i^h) . \tag{28}$$

The log-odds ratios in Equation 28 do not vary with the matching variable score which implies uniform DIF.

The constrained version of the model given in Equation 27 which implies no DIF is

$$\log(m_{ijk}) = \mu + \lambda_j^V + \sum_{g=1}^{d_1} \beta_{1gj} s_k^g + \sum_{h=1}^{d_2} \beta_{2h} r_i^h + \sum_{g=1}^{d_1} \sum_{h=1}^{d_2} \gamma_{gh} s_k^g r_i^h . \tag{29}$$

The model in Equation 29 differs from the model in Equation 27 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 29 and 27 is $d_2(J-1)$. The log-odds ratios in Equation 7 for the model in Equation 29 are all zero implying no DIF.

The models in Equations 25 and 27 can be used to test for nonuniform DIF (the test for uniform DIF has $d_3(J-1)$ degrees of freedom). The models in Equations 27 and 29 can be used to test for uniform DIF (the test for uniform DIF has $d_2(J-1)$ degrees of freedom).

Note that the log-odds in Equations 21 and 23 for the LDFA model are of the same form as the corresponding log-odds in Equations 26 and 28 for the polynomial loglinear model when $d_2 = 1$ and $d_3 = 1$ with the only nonzero γ_{gh} being γ_{11} (in this case the value of d_1 does not affect the log-odds). While the form of these log-odds ratios are the same in this case, the models are different (and not nested), and the estimated cell counts for the two models will not be the same. Even when the log-odds in Equations 26 and 28 have the same parametric form as the log-odds in Equations

21 and 23, the parameter estimates will differ since the models are different. Consequently, the log-odds functions, while both linear, will differ for the polynomial loglinear and the LDFA models.

Choosing a Model

Using the models in Equations 25, 27 and 29 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 25, and are used for the models in Equations 27 and 29 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 25 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 null hypotheses 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 $G_{i-1}^2 - G_i^2 > C$ (if $G_{i-1}^2 - G_i^2 \leq C$ for all i let $i' = 1$). 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 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 example presented later uses a modification of the Haberman procedure to select a model.

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.

Matching Variable

A typical matching variable is 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 that excludes the studied item score will be conditionally independent of the studied item score given the latent variable. Under this condition Meredith and Millsap (1992) show that DIF will exist 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 exist when the test score excluding the studied item score is used as the matching variable. Only under very special conditions will including the studied item in the test score 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 existing 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. However, simulation studies conducted by Donoghue, Holland, and Thayer (1993) and Zwick, Donoghue, and Grima (1993) have indicated that for the Mantel-Haenszel procedure this effect is smaller when the studied item is included (in these studies item responses were simulated using the 3-parameter logistic item response model and the partial credit model).

When the item category scores are equally spaced these scores can be taken to be $0, 1, \ldots, I-1$,

where there are 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

$$m^*_{ijk} = \begin{cases} m_{i,j,k-i+1} & i \leq k \leq K + i - 1 \\ 0 & k < i, k > K + i - 1 \,. \end{cases} \tag{30}$$

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 ($m_{2,j,k}$ is the expected number of examinees who obtain a score of 1 on the item, are in group j, and receive a total test score of $k - 1$). From Equation 30, $m^*_{21k} = m_{2,1,k-1}$ 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 his or her test score excluding the item. For $k = 1$, Equation 30 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. Similarly, $m^*_{11k} = m_{11k}$ for $1 \leq k \leq K$. For $k > K$, $m^*_{11k} = 0$ because if an examinee obtains a score of 0 on the item, he or she cannot obtain a test score larger than the maximum score possible on the other items.

Equation 30 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 (by definition the cell count must be zero). 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 constrained to be zero by 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). Given these considerations and the lack of evidence regarding the relative performance of loglinear models (as opposed to the Mantel-Haenszel procedure) in investigating DIF when the item score is included versus excluded from the test score, the matching variable used in the examples in this paper is the test score *excluding* the item score.

Example

An example of using the polynomial loglinear model to study DIF is given using 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 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 matching variable used for each item is the sum of the item scores on the remaining items (the test score excluding the studied item).

The first step in fitting the loglinear models given in Equations 25, 27, and 29 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 for polytomous items and from 1 to 6 for dichotomous items. For dichotomous items the only possible value of d_2 is 1. The five interaction parameters considered for polytomous items 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}. The six interaction parameters considered for dichotomous items were γ_{11}, γ_{21}, γ_{31}, γ_{41}, γ_{51}, and γ_{61}.

For the dichotomously scored items (items 1 through 21) choosing a model involves choosing values of d_1 and d_3 (the only possible value of d_2 is 1). Instead of applying the Haberman procedure to one sequence of nested models, the Haberman procedure was applied twice — once for d_3 and once for d_1. An error rate of .005 was chosen for each of the two separate Haberman procedures resulting in in an overall error rate of at most .01 (by the Bonferroni inequality) for the two procedures taken as a whole. When selecting d_3, d_1 was set equal to the maximum value (6). The Haberman procedure was applied to a sequence of models given by $d_3 = 1, 2, \ldots, 6$. The overall level of significance chosen was .005, so the value of p^* used for each individual test of the nested models was $1 - (1 - .005)^{1/5} = .001$.

When selecting d_1, d_3 was set equal to the value determined in the first step. The Haberman procedure was applied for the sequence of models corresponding to $d_1 = d_3, d_3 + 1, \ldots, 6$. The overall level of significance was chosen to be .05, so the value of p^* used for each individual test of the nested models was $1 - (1 - .005)^{1/5} = .001$ (there are a maximum of 6 models).

For the polytomous items (items 22 through 27) values must be chosen 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 dichotomous item). A nested sequence of six models were compared for d_3 and d_1. The first six lines gives results for d_3, and the next six lines gives the results for d_1. Chi-square statistics for the models and their degrees of freedom and p-values are presented in the three columns under the heading "Model." Chi-square statistics for comparing adjacent models and their degrees of freedom and p-values are presented in the three columns under the heading "Comparison of Model to Previous Model." For d_3 the first two models to be compared are those given in the first two rows. For both these models $d_1 = 6$ and $d_2 = 1$. The model in the first row has $d_3 = 6$ and the model in the second row has $d_3 = 5$. The chi-square statistic for testing the null hypothesis that the model with $d_3 = 5$ holds against the alternative hypothesis that the model with $d_3 = 6$ holds is given as 1.1890. The value of p^* chosen for each of the tests of consecutive models is .001. Consequently, for the first test ($d_1 = 5$ versus $d_1 = 6$) the null hypothesis of the simpler model is not rejected. None of the tests is significant at the .001 level, so a value of $d_3 = 1$ is chosen (this is indicated in the table by the value of p^* being next to the model with $d_3 = 1$). The next six models correspond to values of d_1 from 1 to 6 (with $d_2 = d_3 = 1$). The first test that is significant at the .001 level is the test for $d_1 = 2$ versus $d_1 = 3$. Consequently, the model selected is $d_1 = 3, d_2 = d_3 = 1$.

For item 23 separate selection procedures were used for d_3, d_2, and d_1. For item 23 the first five lines in Table 1 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 used 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 21.3489 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 $d_2 = 3$ 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 the dichotomous items, d_3 was selected to be equal to 1, and d_1 was selected to be equal to 4 for all items except items 7 and 13, where d_1 was selected to be equal to 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 25, 27, and 29 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 nonuniform DIF, two items showed 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 out the effects against each group and this item exhibited no uniform DIF.

The logistic discriminant function analysis (LDFA) results using the models in Equations 20, 22 and 24 are presented in Table 4. The results in Table 4 differ slightly from the results in Table 3 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. In addition, the analysis reported here used one fewer observation. Still, the chi-square values in Table 4 and the chi-square values in Table 3 of Miller and Spray (1993) are quite similar and there is no pattern in the direction of the differences (in some cases the chi-square in Table 4 is higher and in some cases the chi-square given in Table 3 of Miller and Spray is higher). For these data the effect of including versus excluding the studied item in the matching variable does not seem to be large for the LDFA model.

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

For some of the dichotomous items the log of the odds ratios in Equation 6 were graphed as a function of test score (excluding the studied item) to explore the DIF trends across score levels. Log-odds for the observed counts and fitted counts for the polynomial loglinear and LDFA models for nonuniform DIF were compared. In the graphs of the log of the odds ratios in Equation 6 group 1 is females, group 2 is males, item response category 1 is a correct response, and item response category 2 is an incorrect response. Log-odds ratios above zero indicate that the odds of a correct response was greater for females than for males, and log-odds ratios below zero indicate the odds of a correct response was greater for males than females. Figures 1 through 6 contain odds-ratios graphs for dichotomous items in which significant nonuniform DIF was indicated for the LDFA model but not for the polynomial loglinear model (items 16, 4, 5, 6, 8, 20).

The log-odds in Figures 1 through 6 will be linear functions of the test score for both the LDFA model and the polynomial loglinear model. The log-odd function for the LDFA and polynomial loglinear models for the case of nonuniform DIF are given in Equations 21 and 26, respectively. It can be seen from Equation 21 that the log-odds for the LDFA model are a linear function of the matching variable score. The value of d_3 was set to one for the polynomial loglinear models used for all the dichotomous items (γ_{11j} were the only nonzero γ_{ghj}). In this case Equation 26 is a linear function of the matching variable score.

The graph of the log-odds ratio as a function of test score (excluding the studied item) is presented in Figure 1 for item 16. In Figure 1, the observed log-odds lie primarily above the line indicating no DIF, which is the horizontal line at 0.0. If an item had no DIF, the observed data would approximate this line. The fitted log-odds ratios for the LDFA model have a larger slope than the fitted log-odds ratios for the polynomial loglinear model. The slope of the fitted log-odds for the polynomial loglinear model is small, consistent with the fact that significant nonuniform DIF was not indicated for this 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 these items the slope of the fitted log-odds is larger for the LDFA model than for the polynomial loglinear 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 log-odds ratios for items 15 and 7 are given in Figures 7 and 8. For these items the slope of the fitted log-odds is larger for the polynomial loglinear model than for the LDFA model.

The test for nonuniform DIF was significant for the last 4 polytomous items (items 24 through 27) using the LDFA model, whereas only for item 26 was the test for nonuniform DIF significant using the polynomial loglinear model. For the polytomous items there would be multiple log-odds ratio plots for each item (one plot for each pair of adjacent item response categories) 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 items (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 from the polynomial loglinear model of uniform DIF as a function of test score (excluding the studied item) 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 test 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. For matching variable scores above around 19 the means for males are higher than the means for females in Figure 12, 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.

Examining Common Items for DIF

In the common-item equating design each of the forms of a test used in equating contains a set of items common to all forms. The items common to all forms are called common items, and the items on each form that are unique to that form are called non-common items. The common items may be included with the non-common items 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). The forms are administered to different groups of examinees, at possibly different times (for example, forms administered in different years). The groups receiving the various test forms can be randomly equivalent (the common-item random groups equating design) or chosen in such a way that they are not randomly equivalent (the common-item nonequivalent groups equating design). The result of an equating is a conversion from the scores on the new form to the scores on the old form. If there is more than one new or old form, multiple equatings of pairs of forms are performed. For more information on the common-item equating design see Kolen and Brennan (1995).

For common-item equating to provide valid results it is important that the common items function the same in both the new and old test forms. A common item could function differently on two forms due to the different contexts in which it was embedded (different non-common items), 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 the 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 are those who took, on separate test dates, one of the two forms in which the common item set is embedded.

While both common and non-common items are used in equating, the non-common items are not used in performing a DIF analysis of the common items. Typically, the sum of the common-item scores would be used as the matching variable in a DIF analysis of the common items.

When the groups taking the two forms are not randomly equivalent an item may exhibit DIF solely due to the differences between the two groups (the item exhibits traditional DIF for the two groups). Instead of the item functioning differently due to the different administrations (embedded in different forms and given on different dates), the item is functioning differently due to differences

among the groups. If a common item exhibits DIF there is no way of knowing to what extent the DIF is due to the different groups or the different administrations. Groups taking different forms will typically not be greatly disparate, and will likely be less disparate than groups examined in traditional DIF analysis, so it may be reasonable to attribute DIF observed for a common item to a difference in how the item is functioning on the two administrations, rather than a difference in how the item is functioning in the groups taking the two forms.

Example

An example will be presented of applying the polynomial loglinear model to investigate DIF for common items in a common-item nonequivalent groups equating design. The data used were from a 150 item multiple choice test (all items were dichotomous). 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 also present in the 1993 form) and the 1991 form (administered in 1991 with 38 internal common items also present in 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 statistic for testing the null hypothesis that the nonuniform DIF model with $d_1 = 4$ holds versus the alternative hypothesis that the saturated model holds (which is a goodness of fit test of the nonuniform DIF model) was not significant at the .05 level of significance. This indicates that using $d_1 = 4$ provides an adequate fit to the data.

Three different significance levels were used for the analysis — 0.05, 0.01, and 0.05 / 38 = 0.0013 (a Bonferroni adjustment). The results for all items are presented in Table 5. 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. For nonuniform DIF, only two items were significant at the 0.01 level. Both items that exhibited nonuniform 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 "...not..." (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 than 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 not caught by test development staff who checked for the items being identical on the two forms.

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 the balance in the range of content of the common items should be as similar 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 that exhibited the large amount of uniform DIF, and had syntactic problems, came from content areas with larger numbers of items. Consequently, the items could be removed and no harm would be done to the balance of content in the set of common items.

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; Kolen and Brennan, 1995). 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). 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 statistic for testing the null hypothesis that the nonuniform DIF model with $d_1 = 4$ holds versus the alternative hypothesis that the saturated model holds (which is a goodness of fit test for the nonuniform DIF model) was not significant at the .05 level of significance. This indicates that using $d_1 = 4$ provides an adequate fit to the data.

Again, three different significance levels were used for the analysis — 0.05, 0.01, and 0.05 / 37 = 0.0014 (a Bonferroni adjustment). The results for the 1993/1992 equating are presented in Table 6. 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 nonuniform DIF, only two items were significant at the 0.05 level. Neither of these items exhibited significant uniform DIF.

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 could be less bias when the time interval between administrations is 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 was introduced which incorporated scores for the matching variable and item response categories. 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 accommodates 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 26 versus Equations 15 and 21), 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.

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. The absolute and relative performance of the methods could be studied using simulated data.

The use of DIF techniques for studying whether common items in the common-item equating design function differently on different test forms was discussed. 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 administered. It is important in common-item equating 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. An example was given using the polynomial loglinear model to study DIF in common equating items. In the example presented, two items for which the test for uniform DIF was significant were found to have syntactic differences between the forms.

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 in this paper. The usefulness of confidence bands is demonstrated in Miller and Spray (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.

Donoghue, J. R., Holland, P. W., & Thayer, D. T. (1993). A Monte Carlo study of factors that affect the Mantel-Haenszel and standardization measures of differential item functioning. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 137-166). Hillsdale, NJ: Erlbaum.

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. (1996). Testing for differences in test score distributions using loglinear models. *Applied Measurement in Education, 9*, 305-321.

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.

Kolen, M. J. & Brennan R. L. (1995). *Test equating methods and practices*. New York: Springer-Verlag.

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.

Zwick, R., Donoghue, J. R., & Grima, A. (1993). Assessment of differential item functioning for performance tasks. *Journal of Educational Measurement, 30*, 233.251.

Table 1.

Haberman Procedure for Items 7 and 23.

				Model			Comparison of model to previous model			p-value for
	d1	d2	d3	chi-square	d.f.	p-value	chi-square	d.f.	p-value	choosing a model
Item 7										
	6	1	6	138.7686	156	0.8354				
	6	1	5	139.9576	158	0.8458	1.1890	2	0.5519	
	6	1	4	140.1484	160	0.8690	0.1908	2	0.9090	
	6	1	3	142.6550	162	0.8607	2.5066	2	0.2856	
	6	1	2	147.0201	164	0.8251	4.3651	2	0.1128	
	6	1	1	150.1766	166	0.8053	3.1566	2	0.2063	$p^* = 1 - (1 - .005)^{1/5} = .001$
	6	1	1	150.1766	166	0.8053				
	5	1	1	150.7209	168	0.8265	0.5443	2	0.7618	
	4	1	1	151.7422	170	0.8393	1.0213	2	0.6001	
	3	1	1	162.4350	172	0.6877	10.6929	2	0.0048	$p^* = 1 - (1 - .005)^{1/5} = .001$
	2	1	1	189.3293	174	0.2021	26.8943	2	<0.0001	
	1	1	1	2085.8584	176	<0.0001	1896.5291	2	<0.0001	
Item 23										
	6	3	5	299.1114	322	0.8154				$p^* = 1 - (1 - .003)^{1/4} = .00075$
	6	3	4	320.4603	324	0.5451	21.3489	2	<0.0001	
	6	3	3	336.5308	326	0.3321	16.0705	2	0.0003	
	6	3	2	338.0063	328	0.3399	1.4754	2	0.4782	
	6	3	1	344.8563	330	0.2757	6.8500	2	0.0325	
	6	3	5	299.1114	322	0.8154				$p^* = 1 - (1 - .003)^{1/2} = .0015$
	6	2	5	1281.0891	324	<0.0001	981.9777	2	<0.0001	
	6	1	5	1635.3832	326	<0.0001	354.2941	2	<0.0001	
	6	3	5	299.1114	322	0.8154				
	5	3	5	299.4243	324	0.8326	0.3129	2	0.8552	
	4	3	5	299.8021	326	0.8482	0.3778	2	0.8279	$p^* = 1 - (1 - .003)^{1/5} = .0006$
	3	3	5	339.1881	328	0.3235	39.3860	2	0.0000	
	2	3	5	361.7141	330	0.1107	22.5260	2	<0.0001	
	1	3	5	1926.8813	332	<0.0001	1565.1672	2	<0.0001	

Table 2.

Polynomial Loglinear Models Used for Open-Ended Items.

	Number of Parameters		
Item	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

Note: d_1=4, and $d_2 = d_3 = 1$ were used for all dichotomous items (items 1-21), except for items 7 and 13 where $d_1 = 3$.

Table 3.

Polynomial Loglinear Model Results for Miller & Spray Data.

	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	0.00001 ***	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	0.00001 ***	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	0.00001 ***	1	1.223	0.26879
9	1	26.096	0.00001 ***	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)

Table 4.

Logistic Discriminant Function Analysis Results for Miller & Spray Data.

	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	0.00001 ***	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	0.00001 ***	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	0.00001 ***	1	8.276	0.00402 **
9	1	27.397	0.00001 ***	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 Results for the 1993/1991 Equating Data.

	Uniform DIF			Non-Uniform DIF		
	d.f.	chi-square	p <	d.f.	chi-square	p <
Item						
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	0.00001 ***	1	8.596	0.00337 **
30	1	24.799	0.00001 ***	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 Results for the 1993/1992 Equating Data.

	Uniform DIF			Non-Uniform DIF		
	d.f.	Chi-square	p <	d.f.	Chi-square	p <
item						
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)

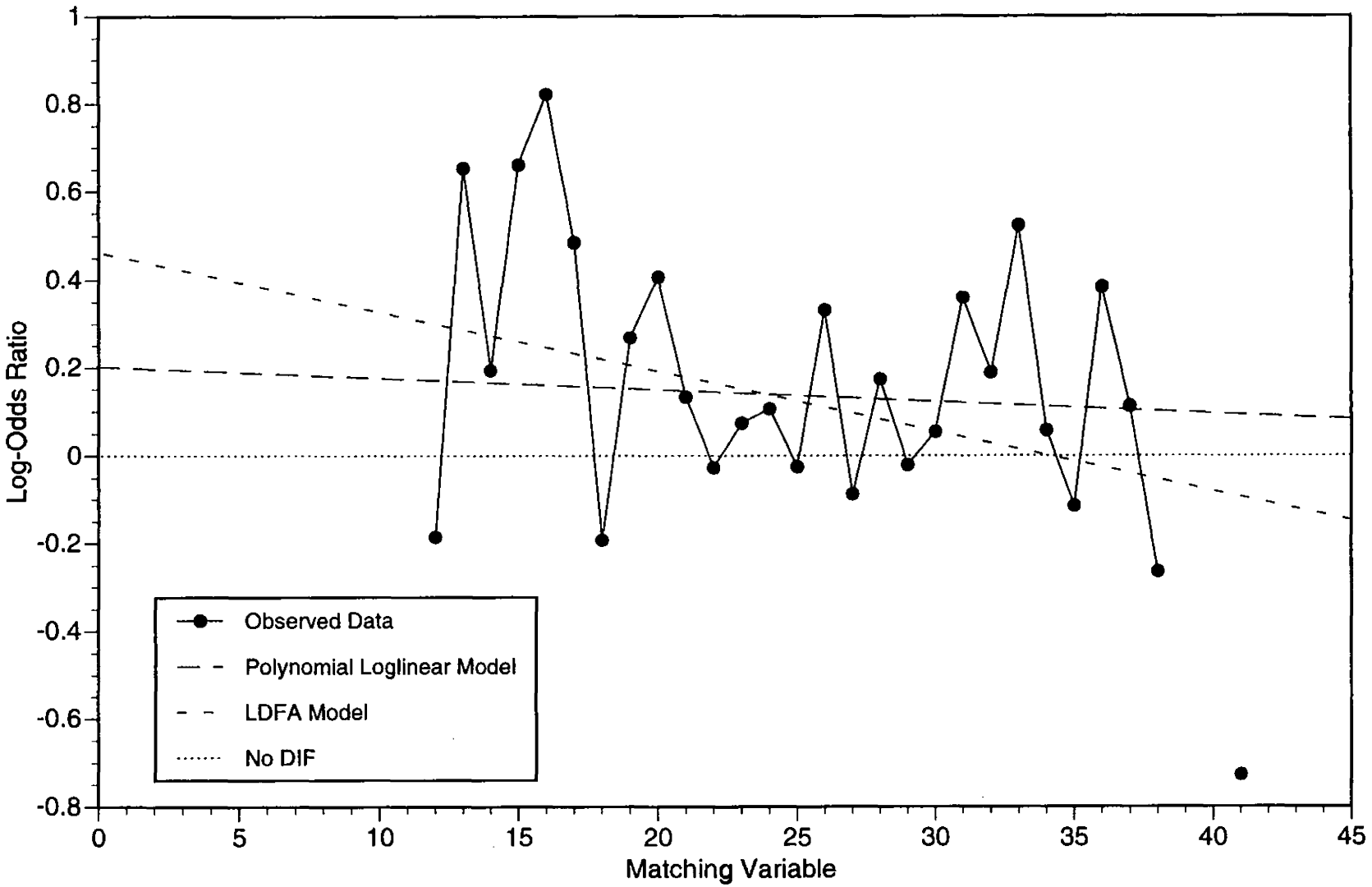

Figure 1. Observed and Fitted Conditional Log-Odds for Item 16.

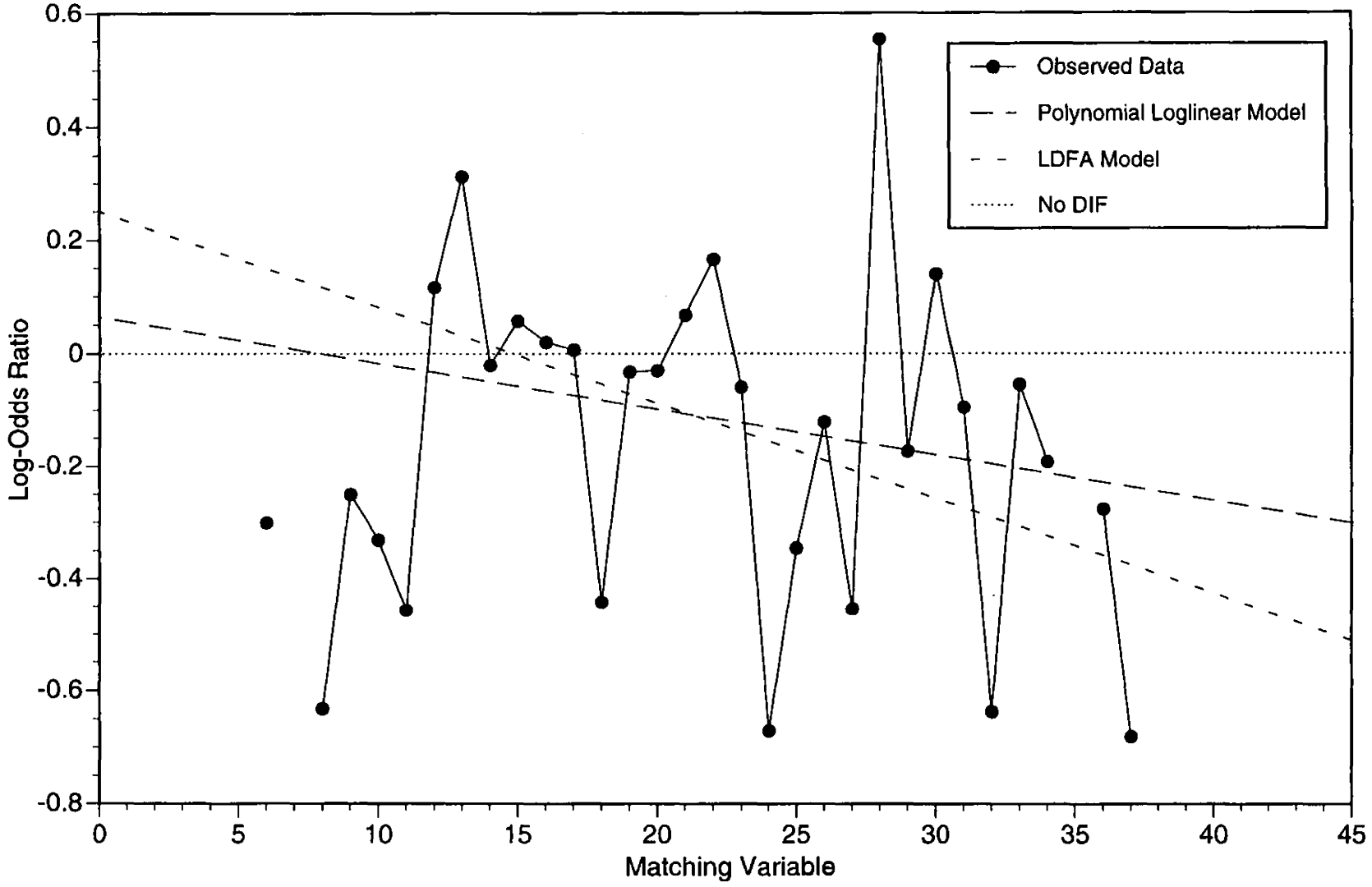

Figure 2. Observed and Fitted Conditional Log-Odds for Item 4.

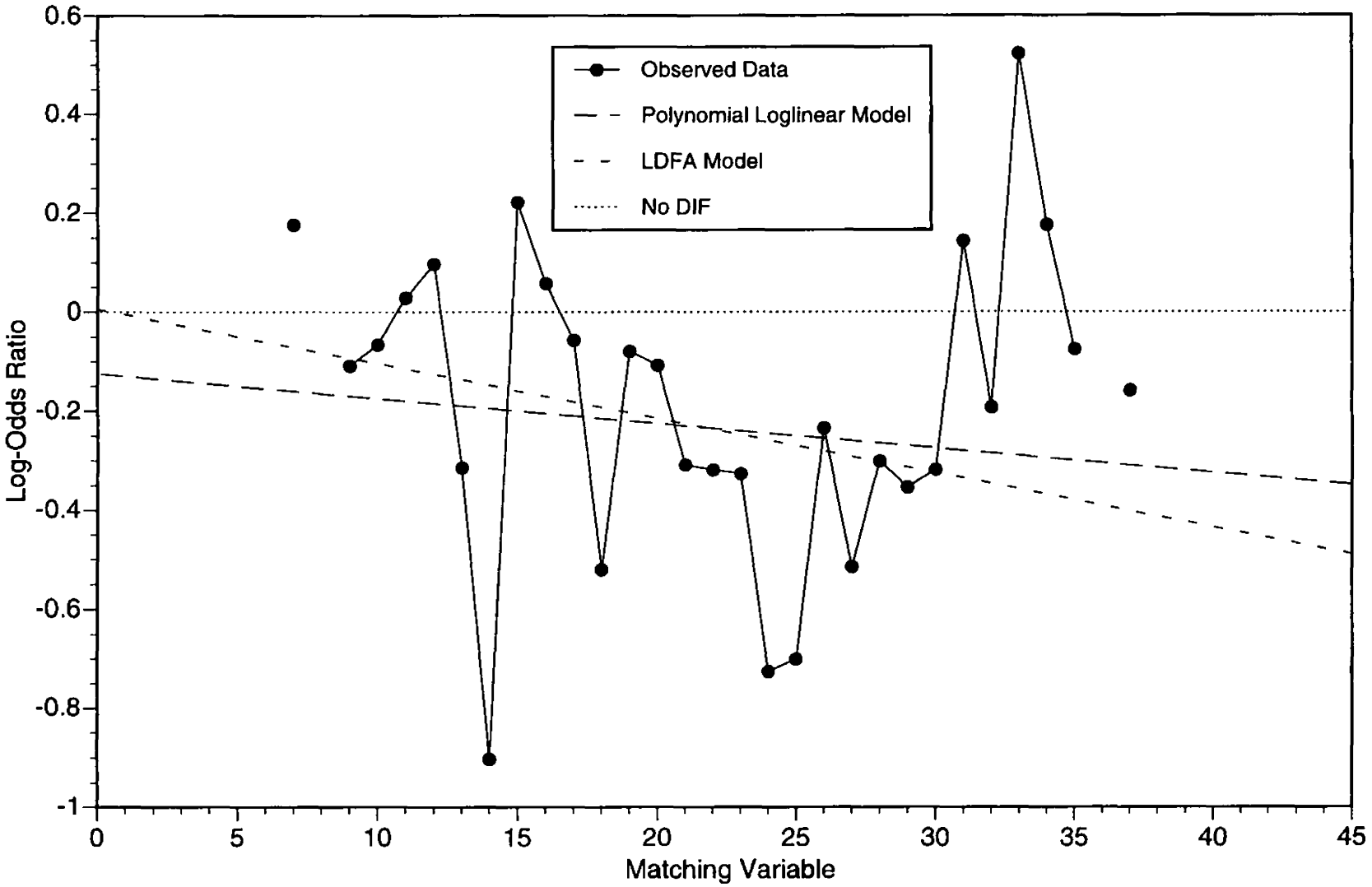

Figure 3. Observed and Fitted Conditional Log-Odds for Item 5.

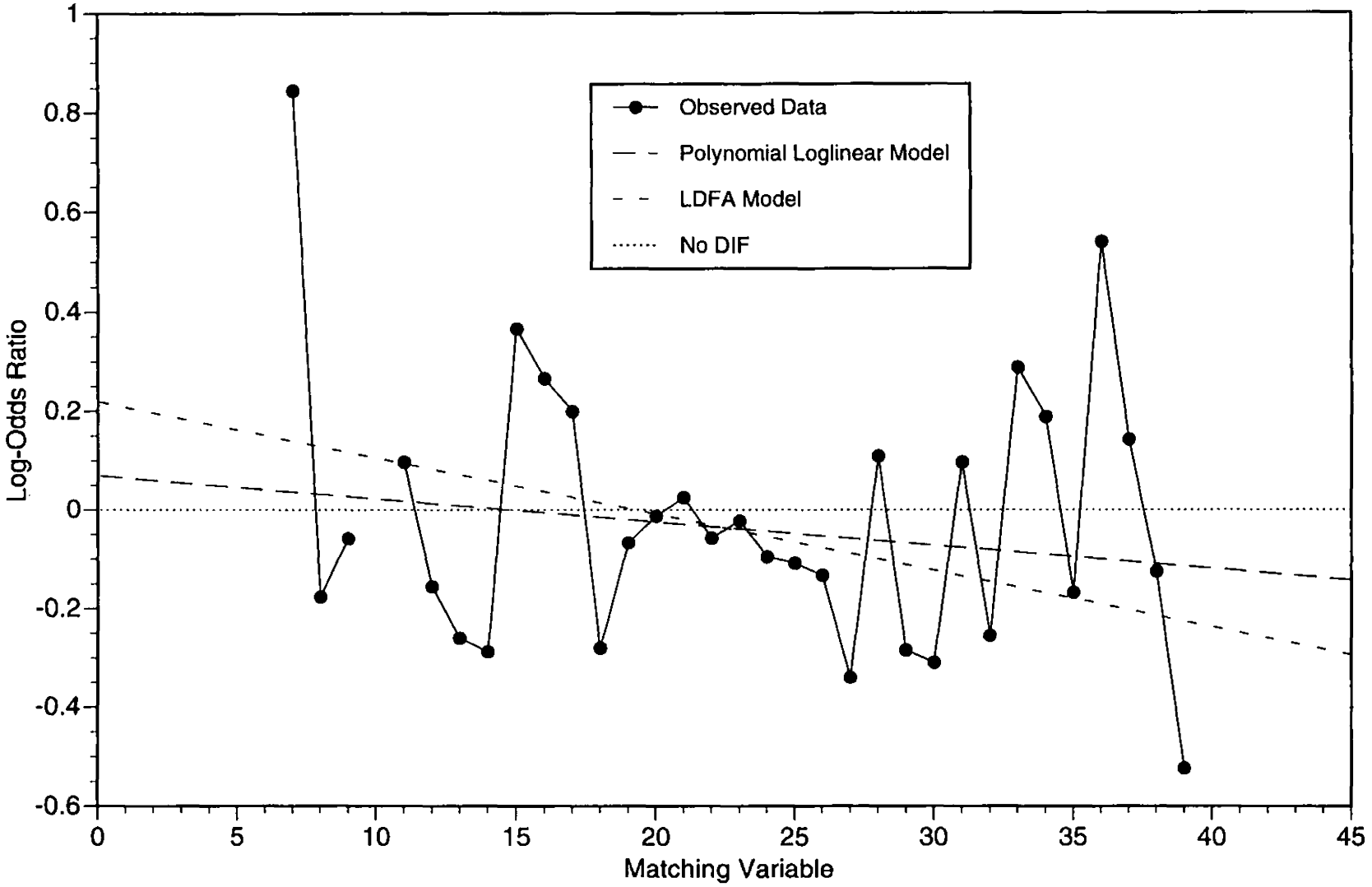

Figure 4. Observed and Fitted Conditional Log-Odds for Item 6.

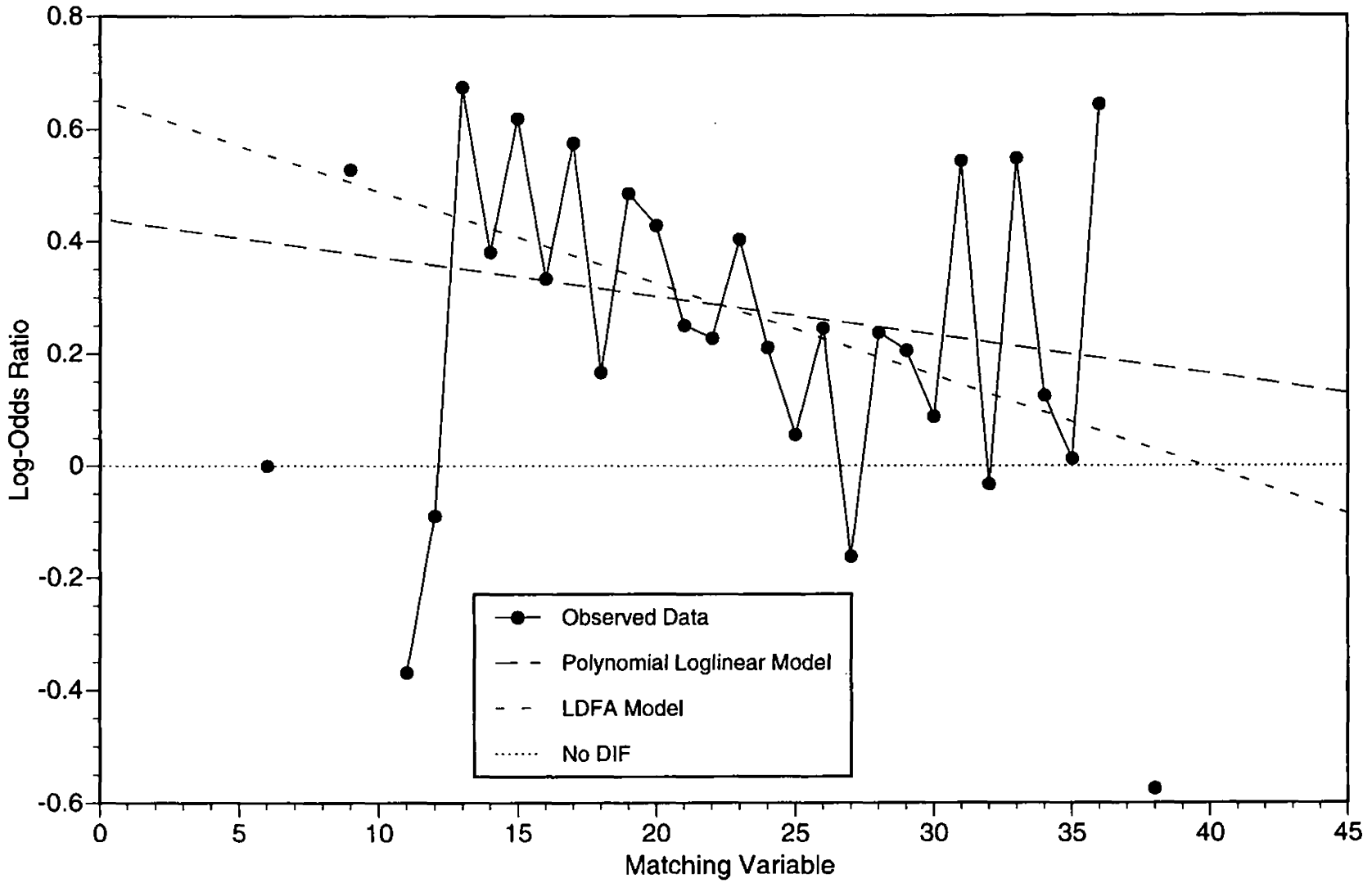

Figure 5. Observed and Fitted Conditional Log-Odds for Item 8.

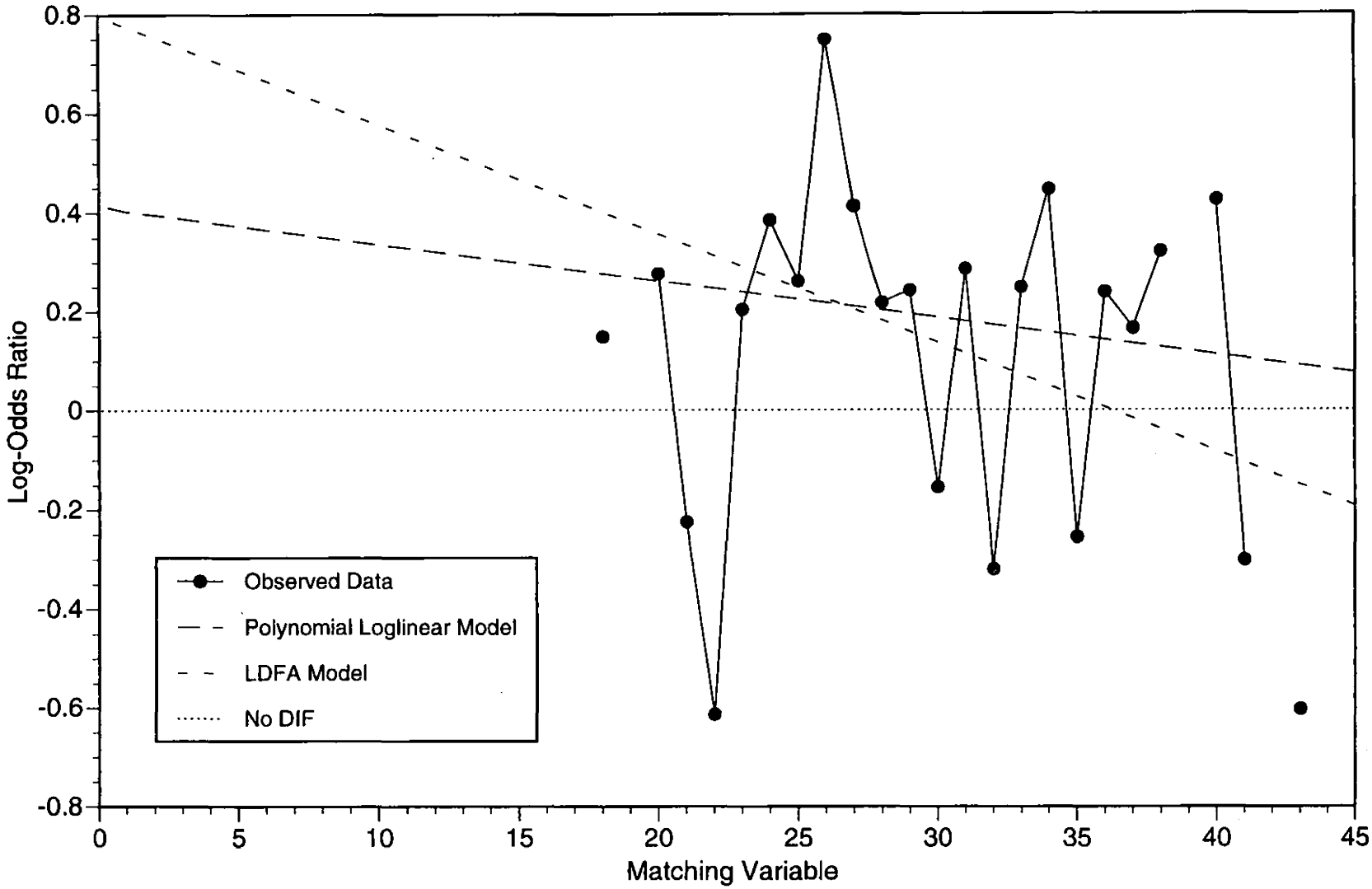

Figure 6. Observed and Fitted Conditional Log-Odds for Item 20.

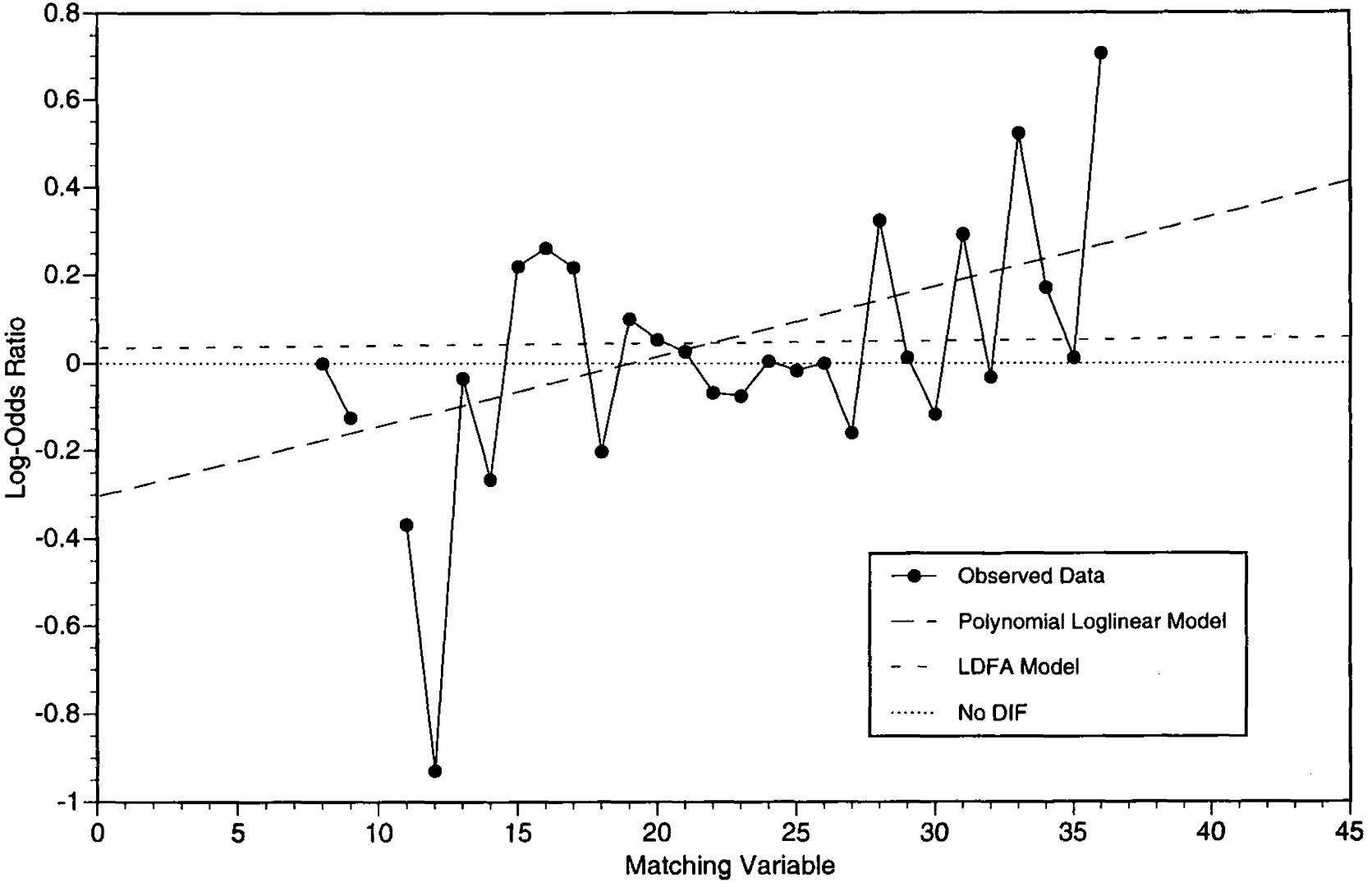

Figure 7. Observed and Fitted Conditional Log-Odds for Item 15.

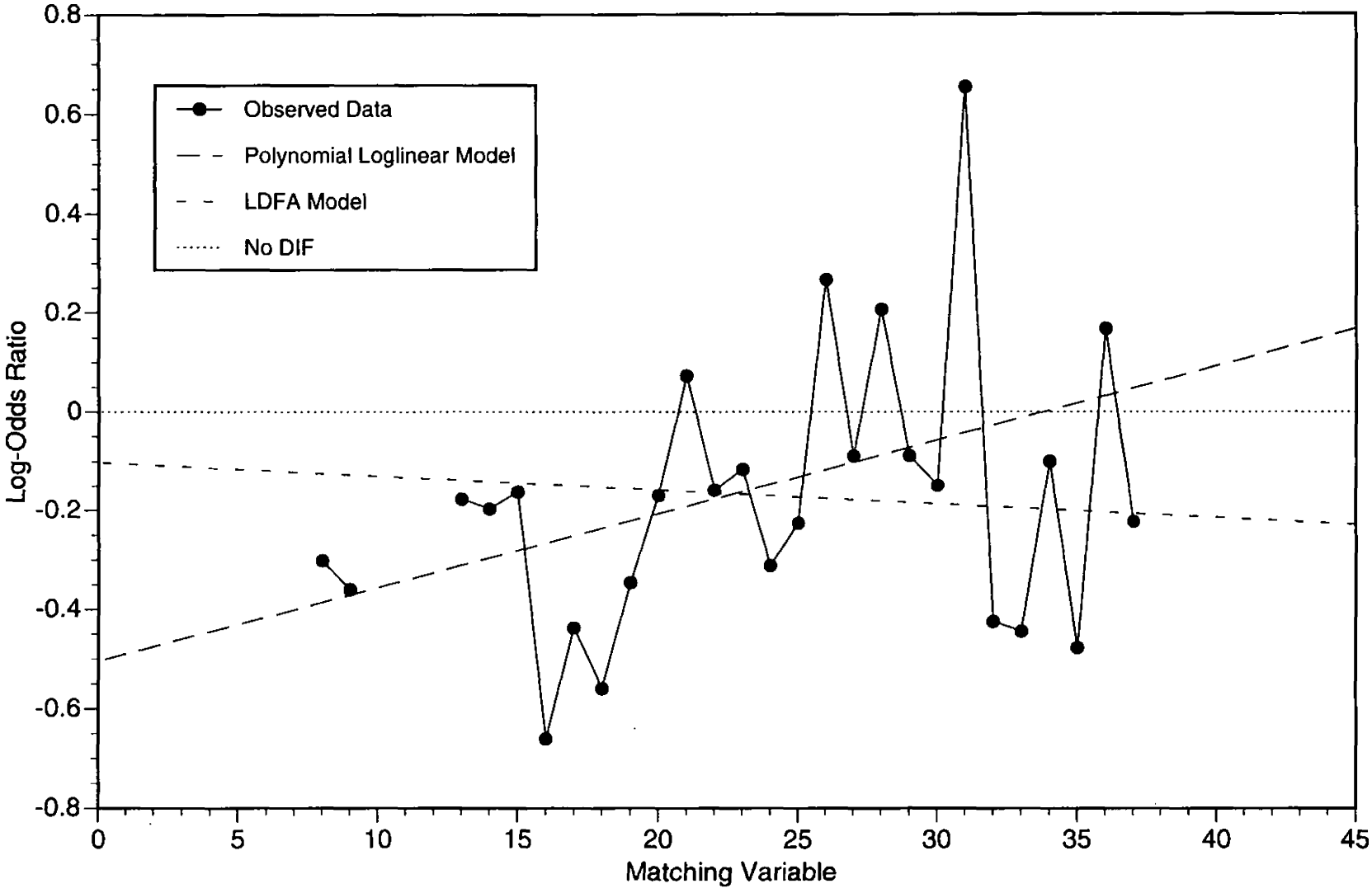

Figure 8. Observed and Fitted Conditional Log-Odds for Item 7.

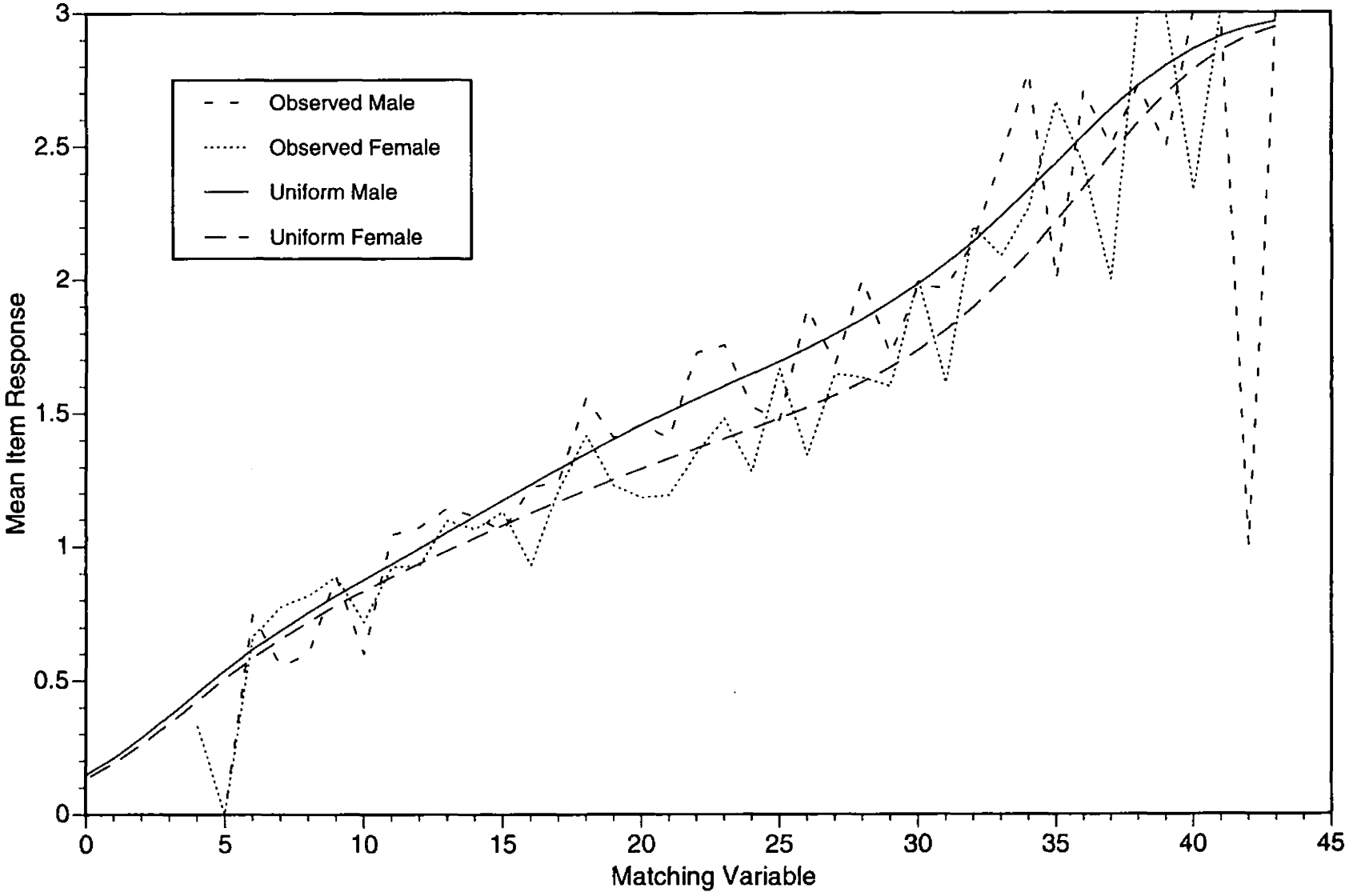

Figure 9. Mean Item Responses from Polynomial Loglinear Model for Uniform DIF Versus Observed Data for Item 23.

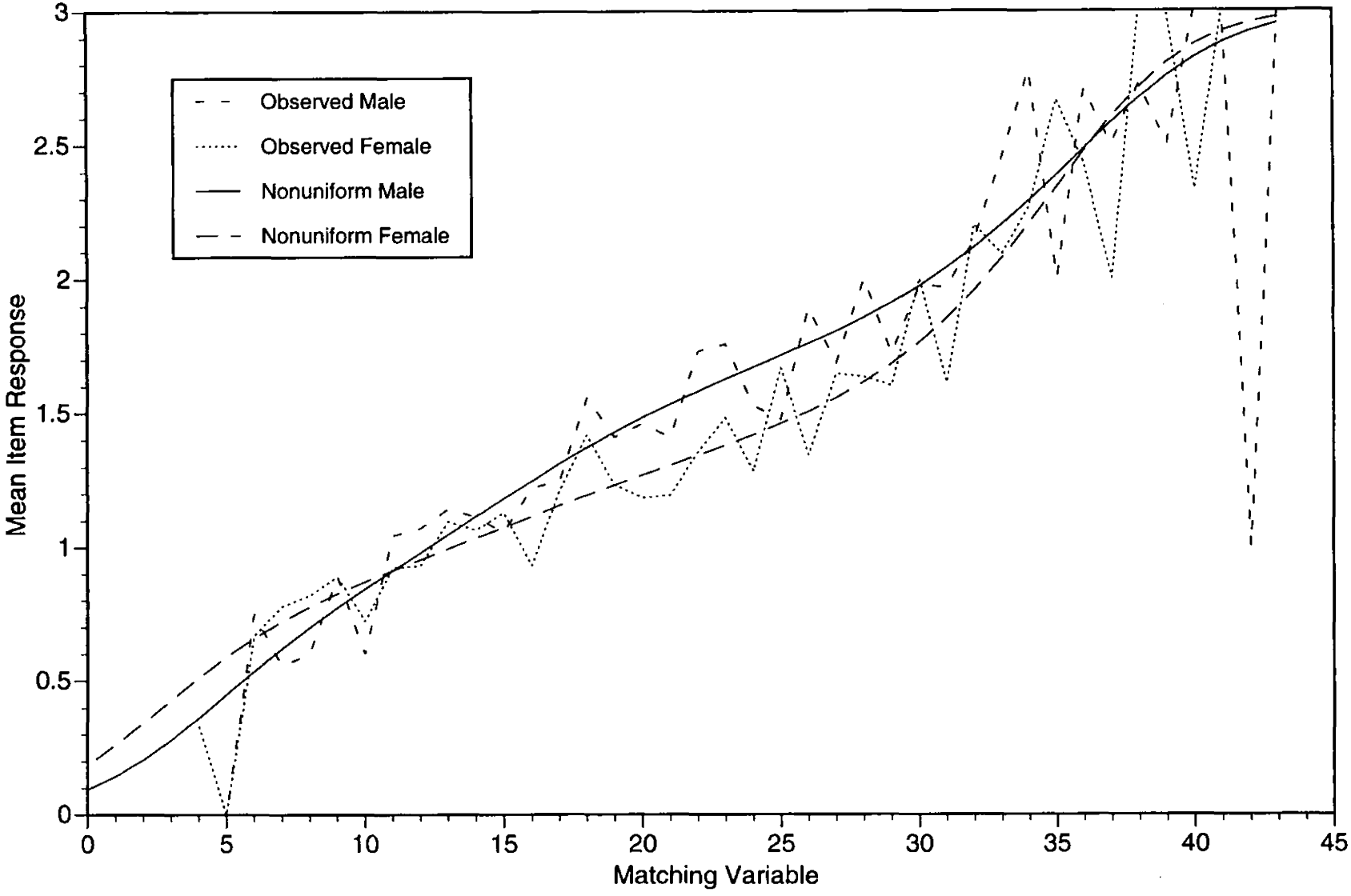

Figure 10. Mean Item Responses from Polynomial Loglinear Model for Nonuniform DIF Versus Observed Data for Item 23.

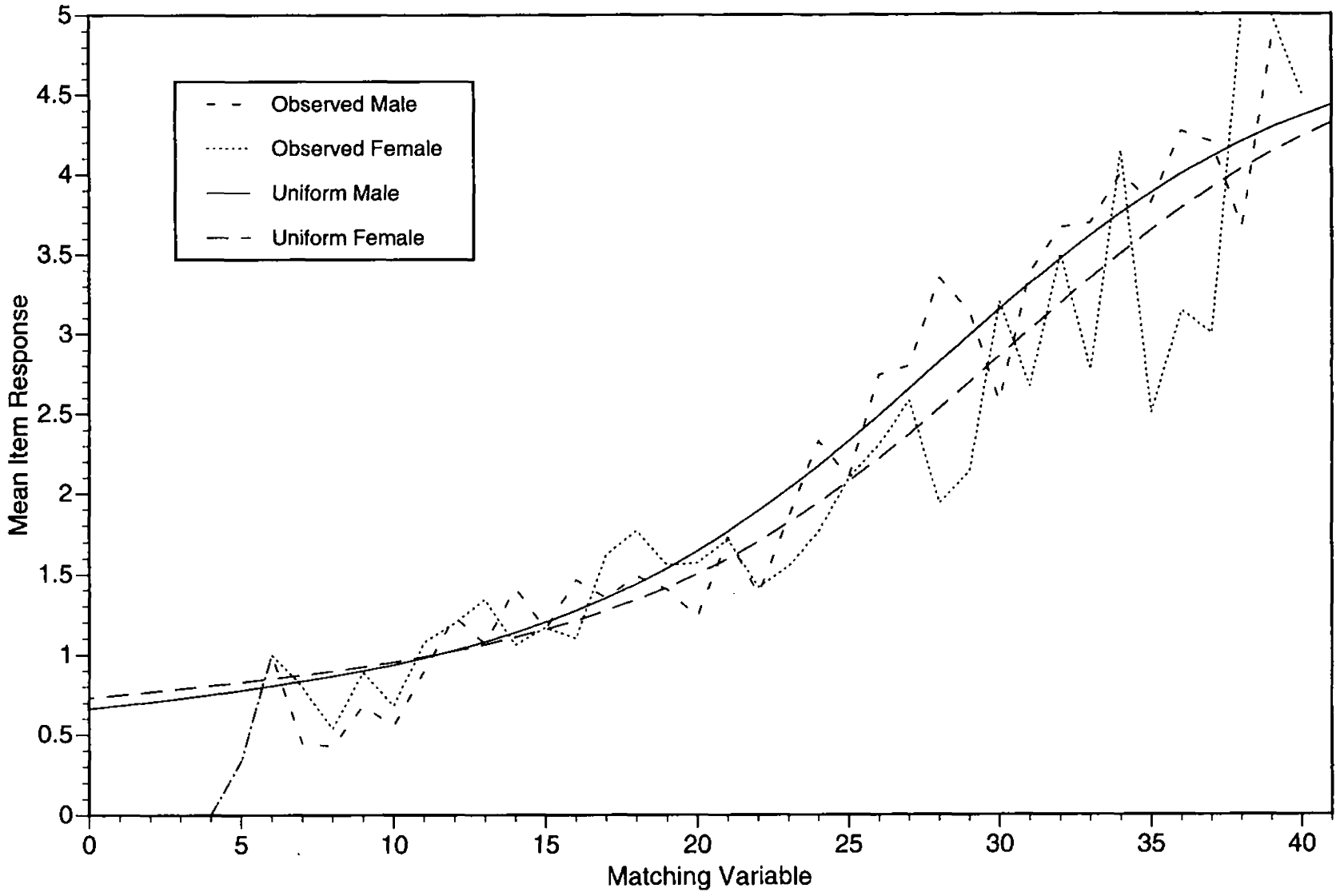

Figure 11. Mean Item Responses from Polynomial Loglinear Model for Uniform DIF Versus Observed Data for Item 26.

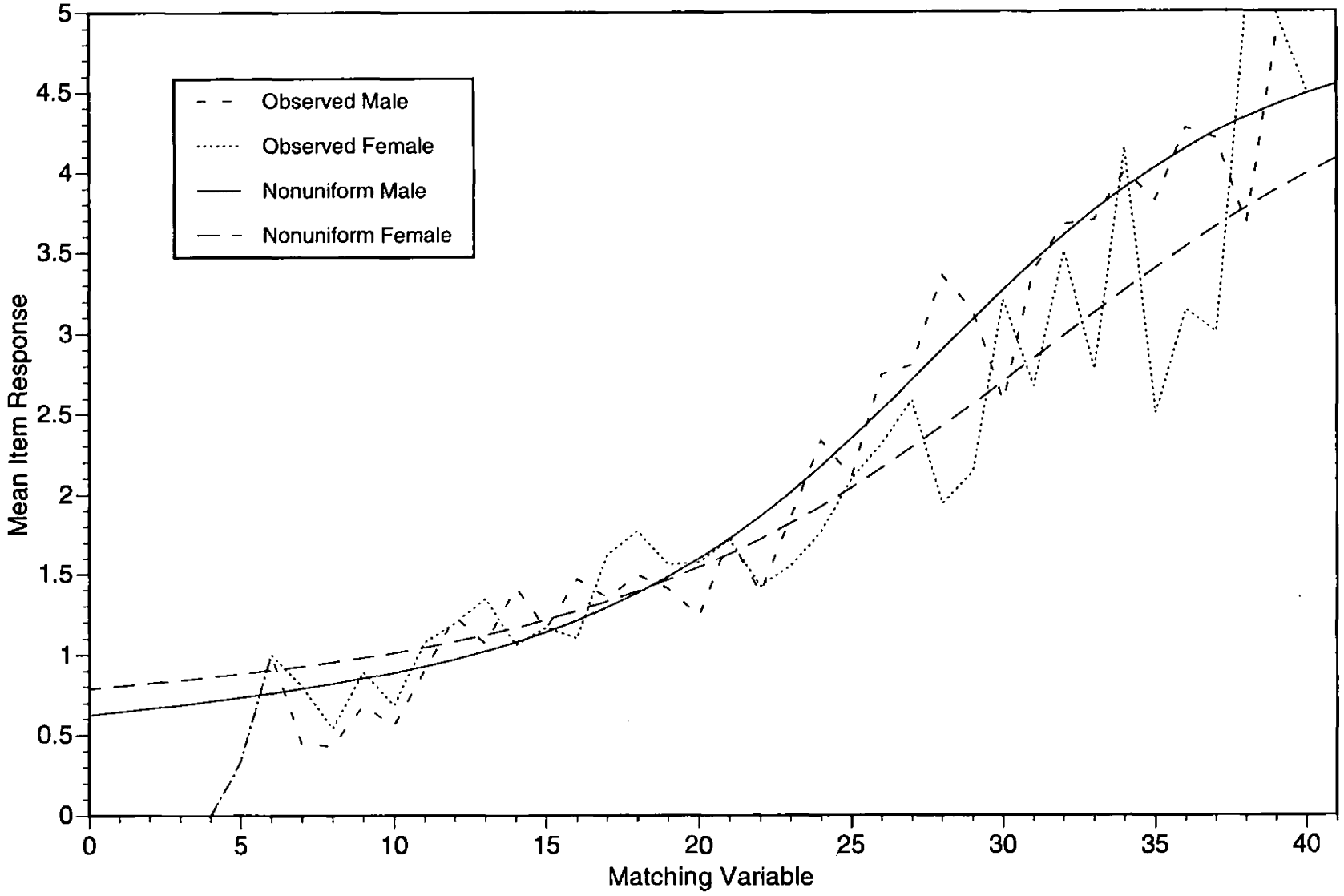

Figure 12. Mean Item Responses from Polynomial Loglinear Model for Nonuniform DIF Versus Observed Data for Item 26.