Using Exact Tests from Algebraic Statistics in Sparse Multi-way Analyses: An Application to Analyzing Differential Item Functioning

Shishir Agrawal[1], Luis David Garcia Puente[2], Minho Kim[2], and Flavia Sancier-Barbosa[2]

[1]Department of Mathematics, University of California San Diego
[2]Department of Mathematics and Computer Science, Colorado College

Abstract

Asymptotic goodness-of-fit methods in contingency table analysis can struggle with sparse data, especially in multi-way tables where it can be infeasible to meet sample size requirements for a robust application of distributional assumptions. However, algebraic statistics provides exact alternatives to these classical asymptotic methods that remain viable even with sparse data. We apply these methods to a context in psychometrics and education research that leads naturally to multi-way contingency tables: the analysis of differential item functioning (DIF). We explain concretely how to apply the exact methods of algebraic statistics to DIF analysis using the R package `algstat`, and we compare their performance to that of classical asymptotic methods.

Keywords: Algebraic statistics, Markov chain Monte Carlo methods, sparse contingency tables, multi-way contingency tables, differential item functioning, item response theory, discrete exponential families, log-linear models, logistic regressions.

1 Introduction

Contingency table analysis has now been an integral part of statistics and the sciences for well over a century (cf. Fienberg & Rinaldo, 2007). Classical methods of analyzing contingency tables, such as Pearson's chi-square test (1900), often involve test statistics that asymptotically follow chi-square distributions as sample size increases. Concerns about the applicability of these methods thus arise in small sample size situations, and it was precisely such concerns that led to the development of Fisher's exact test, which assesses independence in 2×2 tables without relying on large-sample distributional assumptions by

conditioning the test on sufficient statistics for the null model (cf. Fisher, 1922; Fisher & Russell, 1922; Agresti, 2013, section 3.5.1). Fisher's test was later generalized to all two-way tables. Although methods of exhaustive enumeration exist (Mehta & Patel, 1986; Clarkson et al., 1993), they are frequently untenable and Monte Carlo methods are typically more feasible in practice (Boyett, 1979; Patefield, 1981).

In multi-way tables, many statistical models besides joint independence can be of practical interest (e.g., conditional independence, marginal independence, and homogeneous association). As dimensionality increases, it can become infeasible to meet sample size requirements for applying theoretical asymptotics to test the fit of these models. A temptation may arise to alleviate sample size concerns by decreasing dimensionality via marginalization, but this engenders a risk of erroneous inferences such as the Yule-Simpson paradox (cf. Yule, 1903; Simpson, 1951; Lauritzen, 1996, example 4.1). One may resort to collapsing levels, but excessive binning can also be statistically problematic (cf. Deckert & Kummerfeld, 2019, section II.D; Yee & Ho, 2015).

Exact inference via exhaustive enumeration is almost always untenable in multi-way tables. However, pioneering work of Diaconis and Sturmfels (1998) described a Markov Chain Monte Carlo (MCMC) algorithm for exact inference on multi-way tables by drawing on techniques from algebraic and polyhedral geometry. Their methods apply to *discrete exponential families* (cf. Sundberg, 2019; Efron, 2023), which includes a wide range of popular statistical models, including generalized linear models like log-linear models, logistic regressions, and Poisson regressions. Despite early concerns about the computational feasibility of these methods (e.g., Lauritzen, 1996, p. 88), a practical implementation of the methods of Diaconis and Sturmfels now exists in R (R Core Team, 2023) in the comprehensive algebraic statistics package `algstat` (Kahle et al., 2017a), which builds on a range of software packages including `4ti2` (4ti2 Team, 2013) and `LattE` (Baldoni et al., 2013; Kahle et al., 2017b). The field of algebraic statistics has seen significant development since the groundbreaking work of Diaconis and Sturmfels (e.g., Chen et al., 2006; Aoki et al., 2012; Dobra, 2012; Kahle et al., 2018), and further study continues to be warranted as multivariate datasets with small cell counts proliferate.

In order to investigate how the exact methods of algebraic statistics perform against classical asymptotic methods, and to illustrate the use of these methods to practitioners who may have need to analyze sparse multi-way contingency tables, we sought to apply algebraic statistics to a problem in psychometrics and educational research that naturally gives rise to multi-way contingency tables: analyzing *differential item functioning* (DIF, sometimes also called *item bias*). The word *item* in this context refers to a test or survey question, and an item is said to exhibit DIF if there are systematic differences in the distribution of responses from individuals who have the same ability (or some other trait) but who belong to different groups (e.g., different genders, different racial or ethnic groups; cf. Scheuneman, 1979; Mellenbergh, 1982; Hambleton and Swaminathan, 1985, section 13.2). It is typical to treat ability level, group membership, and item response as discrete variables (either categorical or discrete numerical), in which case it becomes necessary to analyze three-way contingency tables. Classical methods of DIF analysis make use of asymptotic goodness-of-fit tests to study these tables. In this study, we propose exact analogs that make use of algebraic statistics. We compare these analogs with their asymptotic counterparts via simulations under various sample sizes and DIF conditions, and demonstrate that these

exact methods can allow researchers to better handle smaller sample sizes.

Using DIF analysis as a case study, we aim broadly to bring to light the existence and performance of the exact methods of algebraic statistics for multi-way analyses. By providing practical examples, we hope that researchers who study DIF, as well as other problems that may give rise to sparse multi-way tables, will consider these exact tests as viable alternatives to classical asymptotic tests.

Our work is organized as follows. In the next section, we give an overview of methods of DIF analysis using two types of statistical models: log-linear models and logistic regressions. We describe classical asymptotic techniques alongside their exact analogs. In the Methods and Results sections, we describe and evaluate simulations that compare the performance of the asymptotic and exact strategies under various sample sizes and DIF conditions. Finally, we discuss our findings and give suggestions for the use of tools from algebraic statistics in DIF analysis and other problems. Code and data related to our work can be found as supplemental material on GitHub.

2   Differential Item Functioning

Differential item functioning (DIF) occurs when individuals from different groups with the same ability (or another trait) have different probabilities of giving the same response to a test or survey item. Three variables are thus involved in DIF analysis: ability, group, and item response. We assume that response is discrete; items with two possible responses are *dichotomous*, while general items are *polytomous*. As is typical, we will compare across two groups — a *reference group* and a *focal group* — though large parts of our discussion apply without difficulty to more than two groups.

An observable variable (e.g., overall test score) is frequently used as a proxy for ability. In order to be able to make use of the exact methods of algebraic statistics for discrete exponential families, we focus on the situation where ability, too, is discretized into finitely many levels. In contexts where ability might naturally be regarded as being continuous, discretization can lead to a loss of information, especially when the number of bins is small (cf. Yee & Ho, 2015; Deckert & Kummerfeld, 2019, section II.D). Nonetheless, binning can also increase parsimony (Deckert & Kummerfeld, 2019, section II.C) and "the (cross-sample) stability of the bias analysis" (van de Vijver & Leung, 2021, p. 70), and it occurs frequently in tests and surveys that classify individuals based on cutoff scores. Researchers choose bins after observing the overall distribution, with no hard rule for determining cutoffs (cf. Scheuneman, 1979, p. 145; van der Flier et al., 1984, pp. 138–9; van de Vijver & Leung, 2021, pp. 69–70; Yee & Ho, 2015, section 5).

A number of parametric formalisms are used for DIF analysis when all three variables of interest are discrete. We focus on two: log-linear models (Mellenbergh, 1982) and logistic regressions (Swaminathan & Rogers, 1990). The two have some differences, and both can be useful in different contexts. Log-linear models may generally be a more flexible option: they require less a priori knowledge about the joint distribution of ability, group, and response, and they can be applied to polytomous items more straightforwardly than logistic regressions (cf. French & Miller, 1996). On the other hand, log-linear models treat ability as a nominal variable; logistic regressions treat it as numerical and can thus better retain numerical or ordinal information about ability levels.

Classical methods of DIF analysis rely on the existence of maximum likelihood estimates (MLEs) and on theoretical asymptotics that hold for large samples. We will explain how algebraic statistics provides exact alternatives that do not make use of either MLE existence or asymptotic distributional assumptions. These exact alternatives may thus present some advantages over asymptotic methods in certain situations. In smaller datasets, for example, it is not uncommon for MLEs to fail to exist due to the presence of zero cell counts. If the number of possible responses or ability levels is very large, it may be infeasible to collect samples that are sufficiently large for a sound application of theoretical asymptotics. The possibility of allowing for a larger number of ability levels may also be useful in situations where ability is binned using cutoff values. For example, Scheuneman (1979, p. 145) presents several considerations for cutoff selection, one of which is avoiding contingency tables with small counts precisely so that asymptotic methods are viable. Replacing asymptotic methods with exact ones can help remove such considerations from cutoff selection, permitting a more principled approach to binning of ability.

2.A Discrete Models for DIF Analysis

Fix a finite set $\mathscr{A}$ of at least two possible ability levels that a subject can have, and a finite set $\mathscr{R}$ of at least two possible responses to the test or survey item under study. Let $\mathscr{G} = \{0, 1\}$ be the set of two groups to which subjects can belong, with 0 indicating the reference group and 1 the focal group. Ability A, group G, and response R are $\mathscr{A}$-, $\mathscr{G}$-, and $\mathscr{R}$-valued random variables, respectively, on the space of respondents. The joint distribution π of the triple (A, G, R) is an unknown distribution on $\mathscr{A} \times \mathscr{G} \times \mathscr{R}$.

We start by specifying a *full model*, i.e., a family $\mathscr{P}$ of distributions on $\mathscr{A} \times \mathscr{G} \times \mathscr{R}$ with no structural zeroes to which π is known a priori to belong. In other words, $\mathscr{P}$ is assumed to lie in the interior of the probability simplex on $\mathscr{A} \times \mathscr{G} \times \mathscr{R}$. Inside $\mathscr{P}$, there are two nested submodels

$$\mathscr{P}_{\mathrm{no23}} \subseteq \mathscr{P}_{\mathrm{no3w}}$$

that are of interest in DIF analysis. The *conditional independence model* $\mathscr{P}_{\mathrm{no23}}$ is the submodel in $\mathscr{P}$ of distributions of (A, G, R) where G and R are conditionally independent given A. The *homogeneous association model* $\mathscr{P}_{\mathrm{no3w}}$ is the submodel in $\mathscr{P}$ of distributions of (A, G, R) where A, G, and R are homogeneously associated (Agresti, 2013, section 9.2.2).[1] The item under study *has no DIF* if $\pi \in \mathscr{P}_{\mathrm{no23}}$. If it does have DIF, it has *uniform DIF* if $\pi \in \mathscr{P}_{\mathrm{no3w}}$ and *nonuniform DIF* otherwise (cf. Mellenbergh, 1982).

Suppose we observe a sample of n subjects in a multinomial sampling scheme, i.e., we make observations of n independent and identically distributed (i.i.d.) random variables with distribution π. We record these observations in a three-way $\#\mathscr{A} \times \#\mathscr{G} \times \#\mathscr{R}$ contingency table, where the cell corresponding to each $(a, g, r) \in \mathscr{A} \times \mathscr{G} \times \mathscr{R}$ contains the number $n_{a,g,r}$ of subjects in the sample with ability a belonging to group g and responding to the item with response r, so that $\sum n_{a,g,r} = n$. Once we settle on strategies for testing the goodness-of-fit of $\mathscr{P}_{\mathrm{no23}}$ and $\mathscr{P}_{\mathrm{no3w}}$ on the observed data, we can organize these tests into the following two-step *DIF analysis paradigm*.

---

[1]The subscript "no23" (resp. "no3w") refers to the fact that $\mathscr{P}_{\mathrm{no23}}$ (resp. $\mathscr{P}_{\mathrm{no3w}}$) is obtained by intersecting $\mathscr{P}$ with the hierarchical log-linear model which does not have an interaction between the *second and third* variables (resp. which does not have a *three-way* interaction between the variables).

(Step 1) **DIF Detection.** Test whether $\mathscr{P}_{\text{no23}}$ fits the observed data. Conclude "no DIF" if it fits. Otherwise, proceed to step 2.

(Step 2) **DIF Classification.** Test whether $\mathscr{P}_{\text{no3w}}$ fits the observed data. Conclude "uniform DIF" if it fits, and "nonuniform DIF" if it doesn't.

2.A.1 Strategies for Testing Goodness-of-Fit

We now describe two possible strategies for testing the goodness-of-fit of the submodels $\mathscr{P}_*$ appearing above. We call these the "asymptotic" and "exact" strategies.

For the asymptotic strategy, we use the observed data to compute a maximum likelihood estimate (MLE) of a distribution in $\mathscr{P}_*$. This MLE can fail to exist in $\mathscr{P}_*$, but if it exists, it is usually unique (cf. Bogdan et al., 2022, lemma 1.1). It may sometimes be computed using a closed-form formula, but the computation usually requires numerical approximation. The divergence of the observed data from the MLE can be measured using the likelihood ratio test statistic

$$G = 2 \sum_{a,g,r} n_{a,g,r} \cdot \ln\left(\frac{n_{a,g,r}}{\hat{\nu}_{a,g,r}}\right), \tag{2.1}$$

where $\hat{\nu}_{a,g,r}$ are the expected counts under the MLE distribution (which are nonzero because the MLE distribution is in $\mathscr{P}_*$ and therefore has no structural zeros).[2] Wilks's theorem (1938; cf. Casella and Berger, 2001, theorem 10.3.3 and section 10.6.2) usually guarantees that the distribution of G as sample size increases is asymptotically chi-square with degrees of freedom equal to the codimension of $\mathscr{P}_*$ inside the probability simplex on $\mathscr{A} \times \mathscr{G} \times \mathscr{R}$. Thus, if the sample size is large enough (a commonly applied heuristic being that at least 80% of expected counts are at least 5), a goodness-of-fit p-value can be computed as the tail area in the asymptotic chi-square distribution beyond the observed value of G.

The asymptotic strategy requires that the MLE for $\mathscr{P}_*$ exists, but we will see that this can often fail. To account for this in our implementations of asymptotic strategies, we say that the conclusion of the analysis is "failure" if the MLE for $\mathscr{P}_{\text{no23}}$ does not exist. Also, if the MLE for $\mathscr{P}_{\text{no23}}$ does exist and DIF is detected at Step 1, but the MLE for $\mathscr{P}_{\text{no3w}}$ does not exist, the analysis concludes that the DIF is present but "unclassifiable."[3]

For the exact strategy, we start by fixing sufficient statistics for n i.i.d. observations from $\mathscr{P}_*$. The *fiber* of the observed table is the (finite) set of all three-way contingency tables with the same value of the sufficient statistics as the observed table. This fiber carries a natural multivariate hypergeometric distribution due to sufficiency of the statistic, and the proportion of tables in the fiber that are at most as likely our observed table is a goodness-of-fit p-value. This is a *conditional p-value* (Agresti, 2013, section 3.5), because we have conditioned on the sufficient statistics.[4]

---

[2] Pearson's chi-square test statistic $\chi^2 = \sum (n_{a,g,r} - \hat{\nu}_{a,g,r})^2 / \hat{\nu}_{a,g,r}$ is a second-order Taylor approximation of G and can be used in its place.

[3] An alternative (and stricter) implementation of the asymptotic strategy might issue "failure" and "unclassifiable" conclusions when heuristic checks for large enough sample size fail.

[4] One can also compute a variant conditional p-value by considering the proportion of tables in the fiber whose deviation from the mean table of the fiber is at least as large as that of our observed table.

That said, the fiber is frequently too large for exhaustive enumeration, so this p-value can rarely be computed in the way just described. Nonetheless, for a number of common choices of $\mathscr{P}$, the algebro-geometric techniques[5] of Diaconis and Sturmfels (1998) can be used to produce a *Markov basis* for the fiber, i.e., a set of moves between tables which connects any two tables in the fiber. This Markov basis can be expensive to compute, but once it has been computed, the Metropolis-Hastings algorithm (Metropolis et al., 1953; Hastings, 1970) can generate a Markov chain that randomly samples the fiber, and the proportion of tables in this Markov chain that are at most as likely as our observed table then serves as our goodness-of-fit p-value. Note that the probability of a table in the fiber depends on the cardinality of the fiber (which is frequently large and difficult to compute), but an *unnormalized probability* can be computed directly from the table without knowing the cardinality of the fiber and these unnormalized probabilities can be compared instead.[6]

2.A.2   *Variants of the DIF Analysis Paradigm*

While we focus on the DIF analysis paradigm as described above, some variants appear in the literature and deserve comment. All of these variants can be implemented using the exact methods of algebraic statistics. One direction of variation (cf. Mellenbergh, 1982; Swaminathan & Rogers, 1990; Dancer et al., 1994) involves switching the order of the goodness-of-fit tests. In other words, we first test the fit of $\mathscr{P}_{\mathrm{no3w}}$, concluding "nonuniform DIF" if it doesn't fit. If it does fit, we then test the fit of $\mathscr{P}_{\mathrm{no23}}$, concluding "uniform DIF" if it doesn't fit and "no DIF" if it does. Our simulations suggest that this "swapped" variant of the paradigm often exhibits worse control over type I error rates for DIF detection, stabilizing noticeably above the nominal rate (cf. the full dataset and the secondary comparison plots on GitHub). It can correspondingly have higher power to detect DIF, but we opted against focusing on it because of its inflated type I error rates.

Another direction of variation (cf. Mellenbergh, 1982; Yesiltas & Paek, 2020) involves augmenting the absolute goodness-of-fit test in the DIF classification step with a relative goodness-of-fit test in which we compare the fit of $\mathscr{P}_{\mathrm{no23}}$ to that of $\mathscr{P}_{\mathrm{no3w}}$, concluding that the item has "uniform DIF" only if $\mathscr{P}_{\mathrm{no3w}}$ fits *and* there is a significant difference between the fit of the two models. If $\mathscr{P}_{\mathrm{no3w}}$ fits but the difference in fit is not significant, the DIF is "unclassifiable." This variation makes the classification step more conservative in the sense that it is less likely to conclude "uniform DIF," but the difference was negligible in our simulations: using the asymptotic strategy, this variant led to a different conclusion in only about 0.1% of simulated tables. We did not include the exact strategy of this variant in our simulations, but a demonstration of conducting model comparisons using algebraic

---

[5]When $\mathscr{P}_*$ is a discrete exponential family, its sufficient statistics parametrize a closed toric subvariety of $\mathrm{Spec}\,k[\mathscr{A} \times \mathscr{G} \times \mathscr{R}]$ over any field k (Diaconis & Sturmfels, 1998, section 3.1; Sturmfels, 1996, chapter 4). The reduced Gröbner basis of the ideal of this affine variety (with respect to any monomial ordering) is a finite set of binomials (Sturmfels, 1996, corollary 4.4), and the exponents that appear in generating binomials correspond to a Markov basis for the fibers of $\mathscr{P}_*$ (Diaconis & Sturmfels, 1998, theorem 3.1). The problem of computing Markov bases thus reduces to the fundamental problem of computational algebraic geometry: computing Gröbner bases. Implicitization (Sturmfels, 1996, algorithm 4.5; Diaconis & Sturmfels, 1998, theorem 3.2; Cox et al., 2015, section 3.3) and integer programming (Sturmfels, 1996, algorithm 12.3; Drton et al., 2009, chapter 1) can be used to compute Gröbner bases of this type.

[6]The ability to use unnormalized probabilities as a stand-in for true probabilities is also important for implementing the Metropolis-Hastings algorithm.

statistics can also be found on GitHub.

A third direction of variation (cf. Yesiltas & Paek, 2020) becomes particularly relevant in situations where one does not have an a priori guarantee that the true distribution lies in the full model $\mathscr{P}$. In this situation, the absolute goodness-of-fit test for $\mathscr{P}_*$ should be replaced with a relative goodness-of-fit test comparing $\mathscr{P}_*$ to $\mathscr{P}$. Our simulations generated data according to distributions which do lie in the choices of $\mathscr{P}$ that we consider, so we did not assess the impact of this direction of variation in this study.

2.B   Introducing the HCI Dataset

To help make our exposition more concrete, we use the `HCI` dataset from the R package `ShinyItemAnalysis` as an example (McFarland et al., 2017; Martinkova & Drabinova, 2018). The dataset concerns the Homeostasis Concept Inventory (HCI), a 20-item multiple-choice instrument that assesses understanding of the biological concept of homeostasis. Each item has four possible responses, one of which is correct; we treat items as dichotomous, with $R = 0$ indicating any incorrect response and $R = 1$ the correct one. Total scores fall in the 18-point range 3–20, and we discretize these into 6 evenly-spaced ability levels $\mathscr{A} = \{0, 1, \ldots, 5\}$. Two possible grouping variables are available in this dataset; here, we consider `major`, with $G = 1$ indicating subjects intending to major in the life sciences.

We focus for the time being on item 17, though we return to the remaining items at the end of this section (cf. table 2.5). Item 17 reads as follows:

> Baroreceptors sense blood pressure. The baroreceptor nerves are cut so the signal from the baroreceptors is unable to reach the cardiovascular control center. After cutting the nerves, blood pressure will
>
> (a) remain constant.
> (b) decrease.
> (c) increase.
> (d) become equal to the set-point value.

The correct answer can be viewed in the supplemental material to McFarland et al. (2017). The three-way contingency table for this item is displayed in table 2.2. Since all observed cell counts are nonzero, all relevant MLEs will exist for this item. The following code snippet constructs this table `t` in R for use further below.

```
library(tidyverse)
library(ShinyItemAnalysis)

# Discretize numerical vector x into discrete levels 0, 1, ..., k-1
discretize <- \(x, k) as.integer(cut_interval(x, k)) - 1

data(HCI)                                          # Load HCI data
t <- as_tibble(HCI) |>
    transmute(ability = discretize(total, 6),      # Discretized ability
              group = major,                       # Group by major
              response = `Item 17`) |>             # Study item 17
    table()                                        # Make three-way table
```

Table 2.2

Three-way contingency table for item 17 of the HCI dataset from the R package `ShinyItemAnalysis` *(McFarland et al., 2017; Martinkova & Drabinova, 2018). Ability is total score discretized into 6 evenly-spaced bins. $R = 1$ indicates a correct response, and $G = 1$ indicates a student majoring the life sciences.*

| | $R = 0$ | | $R = 1$ | |
	$G = 0$	$G = 1$	$G = 0$	$G = 1$
$A = 0$	11	10	4	2
$A = 1$	32	30	10	7
$A = 2$	61	66	27	19
$A = 3$	49	83	14	24
$A = 4$	35	67	14	47
$A = 5$	3	11	5	20

2.C   DIF Analysis with Log-Linear Models

Log-linear models are widely used in categorical data analysis (cf. Agresti, 2013, chapter 9). We concentrate on their application to DIF analysis, as introduced by Scheuneman (1979) and Mellenbergh (1982), who focused on the dichotomous case. The polytomous case was treated more fully later (cf. Dancer et al., 1994; Yesiltas & Paek, 2020).

We take the full model $\mathscr{P}$ to be the *saturated model*, i.e., the set of all distributions on $\mathscr{A} \times \mathscr{G} \times \mathscr{R}$ with no structural zeros. The model $\mathscr{P}$, together with its submodels $\mathscr{P}_{\mathrm{no23}} \subseteq \mathscr{P}_{\mathrm{no3w}}$, can be described as generalized linear models with a logarithmic link function.[7] Namely, for every distribution in $\mathscr{P}$, we can write

$$\ln \Pr[A = a, G = g, R = r] = \lambda + \lambda_a^A + \lambda_g^G + \lambda_r^R + \lambda_{a,g}^{A,G} + \lambda_{a,r}^{A,R} + \lambda_{g,r}^{G,R} + \lambda_{a,g,r}^{A,G,R} \quad (2.3)$$

for all (a, g, r) simultaneously (Agresti, 2013, equation (9.12)), where the λ parameters satisfy appropriate identifiability constraints (cf. Agresti, 2013, section 9.1.4). The submodel $\mathscr{P}_{\mathrm{no3w}}$ is the set of the distributions for which

$$\ln \Pr[A = a, G = g, R = r] = \lambda + \lambda_a^A + \lambda_g^G + \lambda_r^R + \lambda_{a,g}^{A,G} + \lambda_{a,r}^{A,R} + \lambda_{g,r}^{G,R},$$

and $\mathscr{P}_{\mathrm{no23}}$ is the set of the distributions for which

$$\ln \Pr[A = a, G = g, R = r] = \lambda + \lambda_a^A + \lambda_g^G + \lambda_r^R + \lambda_{a,g}^{A,G} + \lambda_{a,r}^{A,R}.$$

These models can also be described as discrete exponential families on $\mathscr{A} \times \mathscr{G} \times \mathscr{R}$ of the form studied by Diaconis and Sturmfels (1998, equation (1.1)).[8] Using a subscript

---

[7]For more about generalized linear models, see, e.g., Agresti (2013, chapter 4), Sundberg (2019, chapter 9), or Efron (2023, chapter 3).

[8]The discrete exponential families of Diaconis and Sturmfels (1998) have trivial *weight function* (cf. Bogdan et al., 2022, section 1.1) and the sufficient statistics of a table can be computed as linear functions of cell counts with nonnegative integer coefficients. To help clarify the translation to the notation of Diaconis and Sturmfels (1998, p. 365), our lists of quantities that form sufficient statistics are specifying the function $t : \mathscr{X}^N \to \mathbb{N}^d \setminus \{0\}$ that appears there with sample size $N = n$ and $\mathscr{X} = \mathscr{A} \times \mathscr{G} \times \mathscr{R}$. For more about exponential families, see, e.g., Sundberg (2019), Bogdan et al. (2022), and Efron (2023).

+ to indicate summing cell counts over all values of that index, the marginals $n_{a,+,r}$ and $n_{a,g,+}$ are sufficient statistics for $\mathscr{P}_{\text{no23}}$. Adding to these the marginals $n_{+,g,r}$ gives sufficient statistics for $\mathscr{P}_{\text{no3w}}$. All of the individual cell counts $n_{a,g,r}$ form sufficient statistics for $\mathscr{P}$.

2.C.1 Asymptotic Goodness-of-Fit Tests

There is a closed-form formula for computing the MLE for $\mathscr{P}_{\text{no23}}$ from the marginals of the observed data; it is the distribution satisfying

$$\Pr[A = a, G = g, R = r] = \frac{n_{a,g,+} n_{a,+,r}}{n_{a,+,+} n},$$

provided that $n_{a,g,+} \neq 0$ and $n_{a,+,r} \neq 0$ for all (a, g, r). The MLE fails to exist if one of these marginals vanishes since we assume no structural zeros (cf. Lauritzen, 1996, corollary 4.9). If the MLE exists, the likelihood ratio test statistic (2.1) asymptotically follows a chi-square distribution with $\#\mathscr{A}(\#\mathscr{G} - 1)(\#\mathscr{R} - 1)$ degrees of freedom. The following R code computes this goodness-of-fit p-value.

```
no23 <- list(c(1, 2), c(1, 3))              # Facet specs of model
x <- loglin(t, no23, fit = TRUE)            # Fit the model
pchisq(x$lrt, x$df, lower.tail = FALSE)     # p-value
```

Note that `loglin` does not check for MLE existence and may run without complaints even when the MLE does not exist. However, for the table `t` from the HCI dataset, this MLE does exist since all observed cell counts are nonzero. Furthermore, 87.5% of the expected counts under the MLE distribution are at least 5, making it reasonable to use an asymptotic chi-square distribution. The p-value is 0.59, suggesting that the item exhibits no DIF.

Since $\#\mathscr{G} = 2$, Eriksson et al. (2006, proposition 12) give a necessary and sufficient characterization for MLE existence for $\mathscr{P}_{\text{no3w}}$: it is necessary that all of the marginals $n_{a,g,+}, n_{a,+,r}$, and $n_{+,g,r}$ be nonzero, but this is not quite sufficient (in the presence of sampling zeroes).[9] There is no closed-form formula for computing the MLE for $\mathscr{P}_{\text{no3w}}$ from the observed data when it exists, but it can be numerically approximated (Agresti, 2013, section 9.7). For example, the R function `loglin` uses iterative proportional fitting (IPF) to compute this MLE (cf. Haberman, 1972). The likelihood ratio test statistic (2.1) asymptotically follows a chi-square distribution with $(\#\mathscr{A} - 1)(\#\mathscr{G} - 1)(\#\mathscr{R} - 1)$ degrees of freedom. Though it is not necessary for DIF analysis of the table `t` of our running example, an asymptotic goodness-of-fit p-value for $\mathscr{P}_{\text{no3w}}$ can be computed by replacing the facet specification `no23` above with `no3w`, defined as follows.

```
no3w <- list(c(1, 2), c(1, 3), c(2, 3))     # Facet specs of model
```

---

[9]The check involves considering all $2 \times 2 \times 2$ collapsings of the table, so it is algorithmically exponential in the dimensions of the table. For other results on MLE existence and uniqueness in discrete exponential families, see Barndorff-Nielsen (2014, theorem 9.13), Jacobson (1989), and Bogdan et al. (2022).

2.C.2   Exact Goodness-of-Fit Tests

Fibers under either model $\mathscr{P}_*$ are frequently far too large to exhaustively enumerate: the function `count_tables` from the R package `algstat` (Kahle et al., 2017a) can be used to compute fiber cardinalities, and for the table `t` of our running example, there are 1,596,426 tables in its fiber under $\mathscr{P}_{\mathrm{no3w}}$, and 103,931,100 under $\mathscr{P}_{\mathrm{no23}}$. However, the techniques of Diaconis and Sturmfels (1998) can be used to compute Markov chains that randomly sample the fibers. The `algstat` function `loglinear` implements the relevant functionality. The following uses the same facet specification `no23` above and computes the exact goodness-of-fit p-value for $\mathscr{P}_{\mathrm{no23}}$ and table `t` of our running example.

```
library(algstat)
x <- loglinear(no23, t)                              # Sample the fiber
x$p.value[["PR"]]                                    # p-value
```

We compute a p-value of 0.60,[10] suggesting again that item 17 exhibits no DIF. Replacing `no23` with `no3w` above gives a p-value for $\mathscr{P}_{\mathrm{no3w}}$ under the exact strategy, although this is again unnecessary for DIF analysis of this example.

2.D   DIF Analysis with Logistic Regressions

Logistic regressions are also widely used in categorical data analysis (cf. Agresti, 2013, chapter 5), and Swaminathan and Rogers (1990) introduced their application to DIF analysis. We assume throughout our discussion of logistic regression that the item under study is dichotomous, i.e., that $\mathscr{R} = \{0, 1\}$. Logistic regression treats ability as a discrete numerical variable. In other words, we also assume that $\mathscr{A}$ is a finite subset of $\mathbb{R}$.

We start by describing the relevant models as generalized linear models with a logit link function. The full model $\mathscr{P}$ is the set of distributions on $\mathscr{A} \times \mathscr{G} \times \mathscr{R}$ for which there exist reals $\tau_0, \tau_1, \tau_2, \tau_3$ such that

$$\mathrm{logit}\,\Pr[R = 1 \mid A = a, G = g] = \tau_0 + \tau_1 a + \tau_2 g + \tau_3 ag, \tag{2.4}$$

for all (a, g), where $\mathrm{logit}(x) = \ln(x/(1-x))$. Note that $\mathscr{P}$ is typically not saturated; it is only saturated when $\#\mathscr{A} = 2$ (and $\#\mathscr{G} = 2$). The submodel $\mathscr{P}_{\mathrm{no3w}}$ is the set of distributions satisfying

$$\mathrm{logit}\,\Pr[R = 1 \mid A = a, G = g] = \tau_0 + \tau_1 a + \tau_2 g,$$

and $\mathscr{P}_{\mathrm{no23}}$ is the set of the distributions satisfying

$$\mathrm{logit}\,\Pr[R = 1 \mid A = a, G = g] = \tau_0 + \tau_1 a.$$

These models can also be described as discrete exponential families. Sufficient statistics for $\mathscr{P}_{\mathrm{no23}}$ are the marginals $n_{a,g,+}$ and $n_{+,+,1}$, together with the quantity

$$\sum_{a \in \mathscr{A}} a \cdot n_{a,+,1}.$$

---

[10] As the Markov chains vary from run to run, so too do the p-values, but this variation is negligible.

Sufficient statistics for $\mathscr{P}_{\text{no3w}}$ are obtained by augmenting the above with the marginal $n_{+,1,1}$. Finally, sufficient statistics for $\mathscr{P}$ are obtained by augmenting all of the above with

$$\sum_{a \in \mathscr{A}} a \cdot n_{a,1,1}.$$

Since the values $a \in \mathscr{A}$ appear in the formulas for the sufficient statistics, we must assume that $\mathscr{A}$ is a finite subset of $\mathbb{N} = \{0, 1, \dots\}$ in order for this model to be of the form considered by Diaconis and Sturmfels (1998). However, this assumption is harmless in practice. Indeed, any real number can be approximated to arbitrary precision by a rational, so we can first assume that every $a \in \mathscr{A}$ is rational, and then we observe that we lose no generality in clearing denominators and translating so that $\mathscr{A} \subseteq \mathbb{N}$. We will therefore make this harmless assumption on $\mathscr{A}$ when discussing logistic regressions.

2.D.1 Asymptotic Goodness-of-Fit Tests

The MLE for the logistic regression model $\mathscr{P}_*$ fails to exist if $n_{a,g,+} = 0$ for some $(a, g) \in \mathscr{A} \times \mathscr{G}$, or if there is (either complete or quasicomplete) separation in the data (Albert & Anderson, 1984, theorems 1, 2(i), and 3). In general, separation can be detected using linear programming (Konis, 2007), as implemented in the R package `detectseparation`. However, for the models of interest to us, more concrete characterizations can be used. For $\mathscr{P}_{\text{no23}}$, separation is equivalent to the existence of $(a^*, r^*) \in \mathscr{A} \times \mathscr{R}$ such that $n_{a,+,r^*} = 0$ for all $a < a^*$ and $n_{a,+,1-r^*} = 0$ for all $a > a^*$. If the MLE exists, the likelihood ratio test statistic (2.1) asymptotically follows a chi-square distribution with $\#\mathscr{A} \cdot \#\mathscr{G} - 2$ degrees of freedom (Agresti, 2013, section 5.2.3). For the table `t` of our running example, the MLE exists since all observed cell counts are nonzero. Moreover, 83.3% of the expected counts under the MLE distribution are at least 5. The R function `glm` uses iteratively reweighted least squares (IRLS) to numerically approximate MLEs for logistic regression. The following code computes the p-value for the goodness-of-fit of $\mathscr{P}_{\text{no23}}$.

```
u <- as_tibble(t) |>
    mutate(across(everything(), as.integer)) |>
    group_by(ability, group) |>
    summarize(p = sum(keep(n, response == 1))/sum(n),
              n = sum(n),
              .groups = "drop")                          # Transform the data
x <- glm(p ~ ability, binomial, weights = n, data = u)   # Fit no23 model
pchisq(x$deviance, x$df.residual, lower.tail = FALSE)    # p-value
```

The p-value is 0.057. With a significance level $\alpha = 0.05$, the resulting DIF analysis suggests that the item has no DIF.

For $\mathscr{P}_{\text{no3w}}$, we can use our assumption that $\mathscr{G} = \{0, 1\}$ to say that separation occurs precisely if there exists $(a_0^*, a_1^*, r^*) \in \mathscr{A} \times \mathscr{A} \times \mathscr{R}$ such that $n_{a,g,r^*} = 0$ for all $(a, g) \in \mathscr{A} \times \mathscr{G} \subseteq \mathbb{R}^2$ strictly to the left of the line ℓ^* that passes through $(a_0^*, 0)$ and $(a_1^*, 1)$, and $n_{a,g,1-r^*} = 0$ for all (a, g) strictly to the right of the same line ℓ^*. If the MLE exists, the likelihood ratio test statistic asymptotically follows a chi-square distribution with $\#\mathscr{A} \cdot \#\mathscr{G} - 3$ degrees of freedom. Replacing the formula `p ~ ability` in the first argument of `glm` above with `p ~ ability + group` gives an asymptotic p-value for the fit of $\mathscr{P}_{\text{no3w}}$ on `t`, though this is again unnecessary for DIF analysis in this case.

2.D.2 Exact Goodness-of-Fit Tests

The R package `algstat` can also be used to compute exact goodness-of-fit p-values for logistic regressions, though, as of this writing, the relevant functionality is not accessible via a single function call (as it is for log-linear models). The following code snippet computes the goodness-of-fit p-value for $\mathscr{P}_{\text{no23}}$ as follows. We first compute a "configuration matrix" `config_no23` which specifies the sufficient statistics for $\mathscr{P}_{\text{no23}}$ in the sense that when the table `t` is regarded as the vector `v`, the product `config_no23 %*% v` is a vector of sufficient statistics of `t`. Then `markov` uses this configuration matrix to compute a Markov basis for the fiber, and `metropolis` runs the Metropolis-Hastings algorithm to sample the fiber.[11]

```
# Algstat C++ function for computing unnormalized probabilities (UProbs)
computeUProbsCpp <- function(x) {
    .Call("_algstat_computeUProbsCpp", PACKAGE = "algstat", x)
}

d <- map(dimnames(t), as.numeric)                   # Dimension names
A <- t(expand_grid(1, d$ability, d$group))          # Auxiliary matrix
v <- as_tibble(t) |>
    arrange(desc(response), ability, group) |>
    pull(n)                                         # Table as a vector
pr <- computeUProbsCpp(matrix(v))                   # UProb of table

config_no23 <- lawrence(A[1:2, ])                   # Configuration matrix
moves <- markov(config_no23, p = "arb")             # Markov basis
sample <- metropolis(v, moves,
                     iter = 10000,
                     burn = 1000,
                     thin = 10)$steps               # Sample fiber
mean(computeUProbsCpp(sample) <= pr)                # p-value
```

For the table `t` of our running example, the exact p-value is 0.04. This suggests that $\mathscr{P}_{\text{no23}}$ does not fit well using a significance level of $\alpha = 0.05$, thus indicating the presence of DIF. To classify the DIF, we compute the p-value for $\mathscr{P}_{\text{no3w}}$ as follows.

```
config_no3w <- lawrence(A)                          # Configuration matrix
moves <- markov(config_no3w, p = "arb")             # Markov basis
sample <- metropolis(v, moves,
                     iter = 10000,
                     burn = 1000,
                     thin = 10)$steps               # Sample fiber
mean(computeUProbsCpp(sample) <= pr)                # p-value
```

The exact p-value is 0.02, suggesting that $\mathscr{P}_{\text{no3w}}$ does not fit well, so we are led to conclude nonuniform DIF. Since we do not have an a priori guarantee that the distribution of the item is in the full logistic regression model $\mathscr{P}$, the poor fit of $\mathscr{P}_{\text{no23}}$ and $\mathscr{P}_{\text{no3w}}$ may

---

[11]There are 939,003,512,241 tables in the fiber of `t` under $\mathscr{P}_{\text{no3w}}$, and 58,866,857,379,038 under $\mathscr{P}_{\text{no23}}$. To emphasize: the former number is almost 1 trillion, and the latter is larger than 58 trillion! This makes evident that methods of exhaustive enumeration are frequently untenable for multi-way tables.

raise doubt as to whether $\mathscr{P}$ fits the data. However, the exact p-value for the goodness-of-fit of $\mathscr{P}$ is 0.12, so the assumption that it fits the data is not unreasonable.

2.E   Revisiting the HCI Dataset

We have seen that for item 17, the asymptotic and exact strategies of DIF analysis agree when using log-linear models, but not when using logistic regressions. The results of applying the same four DIF analyses on each item of the HCI dataset are displayed in table 2.5. For many of the items, the asymptotic and exact strategies agree. When they differ, the pattern displayed by item 17 is somewhat atypical: differences more frequently result from MLE nonexistence. MLEs for the log-linear conditional independence model fail to exist for 6 of the 20 items when total score is binned into 6 ability levels (and this number increases to 15 out of 20 when 9 bins are used instead). The exact strategy using log-linear models is unaffected by MLE nonexistence and concludes no DIF in every case.

Overall, the results in table 2.5 are indicative of a fair test, corroborating the claim that "the HCI assesses understanding of homeostasis for students who are pursuing majors in the life sciences and for students pursuing other majors, although life science majors tend to show the best performance" (McFarland et al., 2017, p. 6).[12]

3   Methods

3.A   Simulation Models

To compare the performance of the various DIF analysis methods discussed above, we simulated data using Bock's nominal response model (1972), which is a nominal poly-tomous generalization of Rasch's dichotomous model.[13] Fix a finite set of ability levels $\mathscr{A} \subseteq \mathbb{N}$, two groups $\mathscr{G} = \{0, 1\}$, and a finite set of responses $\mathscr{R}$. Fix also a joint distribution for (A, G). In order to specify the joint distribution of (A, G, R), it is enough to specify the conditional distribution of R given (A, G), which we do by choosing $u, v \in \mathbb{R}^{\mathscr{G} \times \mathscr{R}}$ and declaring $\Pr[R = r \mid A = a, G = g]$ to be proportional to $\exp(u(g, r)a + v(g, r))$. Explicitly,

$$\Pr[R = r \mid A = a, G = g] = \frac{\exp(u(g, r)a + v(g, r))}{\displaystyle\sum_{r' \in \mathscr{R}} \exp(u(g, r')a + v(g, r'))}. \tag{3.1}$$

Since $\mathscr{G} = \{0, 1\}$, any $x \in \mathbb{R}^{\mathscr{G} \times \mathscr{R}}$ corresponds uniquely to $x_0, x_\Delta \in \mathbb{R}^{\mathscr{R}}$ such that

$$x(g, r) = x_0(r) + x_\Delta(r)g.$$

Decomposing u and v in this way, observe that (u_0, v_0) determines the response distribution in the reference group 0, and (u_Δ, v_Δ) describes how the response distribution changes as we go from the reference group 0 to the focal group 1.

---

[12]The claim that "life science majors tend to show the best performance" can be quantified, for example, by tiny p-values for fit of the model of *marginal* independence of ability and group.

[13]Samejima's graded response model (1968) is an *ordinal* generalization of Rasch's dichotomous model, as opposed to Bock's *nominal* one. We use Bock's model because it is simpler to analyze theoretically.

Table 2.5

*DIF analysis results for each item of the HCI dataset with a significance level of $\alpha = 0.05$.
We include results of binning total score into 6 and 9 ability levels. "Failure" means that
the MLE for $\mathscr{P}_{\mathrm{no23}}$ does not exist, daggers mark results that change to "none" after
BH-adjustments for multiple testing for the fit of $\mathscr{P}_{\mathrm{no23}}$ (Benjamini & Hochberg, 1995),
and stars mark results where the BH-adjusted p-value for the full model $\mathscr{P}$ is small (so the
assumptions of the DIF analysis may not be satisfied). Source code is on GitHub.*

#$\mathscr{A}$	Item	Log-Linear Models		Logistic Regressions	
		Asymptotic	Exact	Asymptotic	Exact
6	1	*failure*	*none*	none	none
	2	none	none	none	none
	3	*failure*	*none*	none	none
	4	none	none	nonuniform[†]	nonuniform[†]
	5	none	none	none	none
	6	none	none	none	none
	7	none	none	none	none
	8	*failure*	*none*	none	none
	9	none	none	none	none
	10	none	none	nonuniform[*]	nonuniform[†]
	11	none	none	none	none
	12	none	none	none	none
	13	none	none	none	none
	14	*failure*	*none*	none	none
	15	none	none	none	none
	16	none	none	none	none
	17	none	none	*none*	*nonuniform*[†]
	18	*failure*	*none*	none	none
	19	*failure*	*none*	none	none
	20	none	none	none	none
9	1	*failure*	*none*	none	none
	2	none	none	none	none
	3	*failure*	*none*	none	none
	4	*failure*	*none*	none	none
	5	*failure*	*none*	none	none
	6	*failure*	*none*	none	none
	7	none	none	none	none
	8	*failure*	*none*	none	none
	9	*failure*	*none*	none	none
	10	*failure*	*none*	nonuniform[†]	nonuniform[†]
	11	*failure*	*none*	none	none
	12	*failure*	*none*	none	none
	13	*failure*	*none*	none	none
	14	*failure*	*none*	none	none
	15	none	none	none	none
	16	none	none	none	none
	17	none	none	nonuniform[†]	nonuniform[†]
	18	*failure*	*none*	none	none
	19	*failure*	*none*	none	none
	20	*failure*	*none*	none	none

Let $C_{\mathscr{R}}$ be the subspace of functions in $\mathbb{R}^{\mathscr{G} \times \mathscr{R}}$ which do not depend on the second coordinate, and identify $C_{\mathscr{R}}$ with $\mathbb{R}^{\mathscr{G}}$. If $u', v' \in C_{\mathscr{R}}$, then

$$\exp((u + u')(g, r)a + (v + v')(g, r)) = \exp(u(g, r)a + v(g, r)) \underbrace{\exp(u'(g)a + v'(g))}_{\text{does not depend on } r},$$

so the underbraced expression factors out of both the numerator and denominator of (3.1). In other words, modifying u and v by vectors in $C_{\mathscr{R}}$ does not change the resulting distribution, so we lose no generality in choosing u and v to lie in some fixed subspace of $\mathbb{R}^{\mathscr{G} \times \mathscr{R}}$ complementary to $C_{\mathscr{R}}$. Since $C_{\mathscr{R}}$ can be described as the space of $x \in \mathbb{R}^{\mathscr{G} \times \mathscr{R}}$ for which x_0 and x_Δ lie in the subspace $C = \text{span}\{1\} \subseteq \mathbb{R}^{\mathscr{R}}$ of constant functions, we can assume that u and v are chosen so that $u_0, u_\Delta, v_0, v_\Delta \in C^\perp$.

3.B Comparison to Log-Linear Models

The vector (u_Δ, v_Δ) determines the qualitative type of DIF in the model. Indeed, observe that

$$\ln \Pr[A = a, G = g, R = r]$$
$$= \ln \Pr[A = a, G = g] \Pr[R = r \mid A = a, G = g]$$
$$= \underbrace{\ln \Pr[A = a, G = g] - \ln \left(\sum_{r' \in \mathscr{R}} \exp(u(g, r')a + v(g, r')) \right)}_{(*)(a,g)} + \ln \exp(u(g, r)a + v(g, r))$$
$$= (*)(a, g) + v_0(r) + u_0(r)a + v_\Delta(r)g + u_\Delta(r)ag.$$

The only terms that exhibit dependence on both group and response are the last two, and only the last exhibits dependence on all three variables. Thus G and R are conditionally independent given A if and only if $u_\Delta = v_\Delta = 0$, and the three variables are homogeneously associated if and only if $u_\Delta = 0$.

Stated differently, the distribution has no DIF if and only if $u_\Delta = v_\Delta = 0$, and it has uniform DIF if and only if $u_\Delta = 0$ but $v_\Delta \neq 0$. In all other cases, the distribution has nonuniform DIF. We call $(u_\Delta, v_\Delta) \in (C^\perp)^2$ the "DIF vector" of the model. See figure 3.2. We use the euclidean norm of (u_Δ, v_Δ) as a measure of the size of the DIF.[14]

3.C Comparison to Logistic Regressions

If $\mathscr{R} = \{0, 1\}$, then equation (3.1) gives

$$\text{logit} \Pr[R = 1 \mid A = a, G = g]$$
$$= u(g, 1)a + v(g, 1) - u(g, 0)a - v(g, 0)$$
$$= \underbrace{(v_0(1) - v_0(0))}_{\tau_0} + \underbrace{(u_0(1) - u_0(0))}_{\tau_1} a + \underbrace{(v_\Delta(1) - v_\Delta(0))}_{\tau_2} g + \underbrace{(u_\Delta(1) - u_\Delta(0))}_{\tau_3} ag.$$

---

[14]It is worth noticing that DIF is a property of the conditional distribution of R given A and G. After $\mathscr{A}$ and $\mathscr{R}$ are fixed, this conditional distribution is determined by $(u_0, u_\Delta, v_0, v_\Delta) \in (C^\perp)^4$, not by $(u_\Delta, v_\Delta) \in (C^\perp)^2$ alone. There exist measures of DIF size that may be more statistically meaningful than the norm of (u_Δ, v_Δ) and which are not functions of that norm alone (cf. the Kullback-Leibler heat map on GitHub).
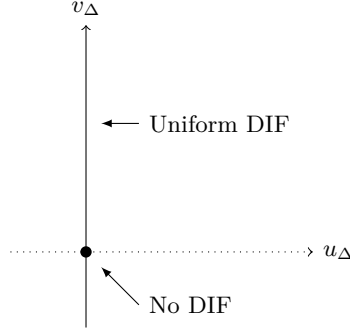
Figure 3.2

The space $(C^\perp)^2$ of all DIF vectors (u_Δ, v_Δ) for a fixed (u_0, v_0). The picture is heuristic since $\dim(C^\perp) = \#\mathscr{R} - 1$ in general. The origin corresponds to no DIF, the remainder of the "vertical axis" $u_\Delta = 0$ corresponds to uniform DIF, and nonuniform DIF occurs generically everywhere else.

In other words, the dichotomous items we simulate using model (3.1) induce distributions on $\mathscr{A} \times \mathscr{G} \times \mathscr{R}$ that lie in the full logistic regression model (2.4).

3.D Simulations and Assessment

We chose 16 integer seeds to create a variety of simulation models. Using each of these seeds, we generate a random choice of $\mathscr{A}$ and $\mathscr{R}$, a joint distribution of (A, G), a vector $(u_0, v_0) \in (C^\perp)^2$, and five DIF vectors $(u_\Delta, v_\Delta) \in (C^\perp)^2$. The five DIF vectors are:

(1) the zero vector,

(2) a uniform unit vector (i.e., with $u_\Delta = 0$),

(3) twice the same uniform vector,

(4) a nonuniform (i.e., generic) unit vector, and

(5) twice the same nonuniform vector.

In other words, each integer seed gives rise to five joint distributions of (A, G, R), all of which have the same distribution of (A, G), and the same conditional distribution of R given A and $G = 0$, but which exhibit five different types and magnitudes of DIF. Model parameters generated by 3 of the 16 integer seeds that we used are displayed in table 3.3. For model parameters for other seeds, see GitHub.

For each of the five joint distributions of (A, G, R) corresponding to each seed, we simulate collecting samples of various sizes. Each sample is handed off to a number of DIF analysis methods: every table is handed off to the asymptotic and exact versions of the method using log-linear models, and every dichotomous table is also handed off to the asymptotic and exact versions of the method using logistic regression. The result of each method can then be compared against the known "true" value of the type of DIF in the simulation model. This allows us to compare the performance of the DIF analysis methods as sample sizes and DIF conditions vary.
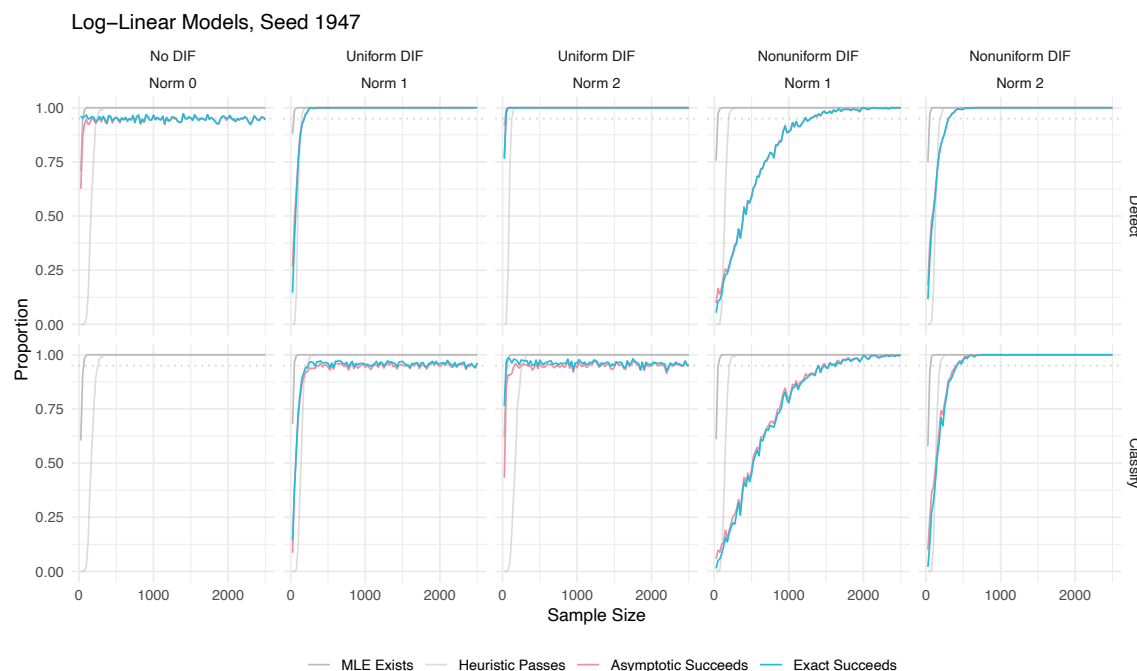
Table 3.3

Simulation model parameters generated by 3 of the 16 integer seeds we chose. In all instances, $\mathscr{G} = \{0, 1\}$. These are nominal response models, so elements of $\mathscr{R}$ are just labels. The displayed DIF vectors (u_Δ, v_Δ) have norm 1 (multiply by 2 for norm 2).

Seed	Parameter	Value
1930	$\Pr[G = 0]$	0.366
	$\mathscr{A}$	{0, 1, 2, 3, 4, 5, 6}
	$\Pr[A \mid G = 0]$	(0.002, 0.020, 0.095, 0.235, 0.328, 0.244, 0.076)
	$\Pr[A \mid G = 1]$	(0.005, 0.043, 0.152, 0.286, 0.302, 0.170, 0.040)
	$\mathscr{R}$	{0, 1}
	(u_0, v_0)	(0.407, -0.407, 0.579, -0.579)
	Uniform (u_Δ, v_Δ)	(0.000, 0.000, 0.707, -0.707)
	Nonuniform (u_Δ, v_Δ)	(0.004, -0.004, 0.707, -0.707)
1947	$\Pr[G = 0]$	0.429
	$\mathscr{A}$	{0, 1}
	$\Pr[A \mid G = 0]$	(0.368, 0.632)
	$\Pr[A \mid G = 1]$	(0.203, 0.797)
	$\mathscr{R}$	{0, 1}
	(u_0, v_0)	(-0.331, 0.331, -0.625, 0.625)
	Uniform (u_Δ, v_Δ)	(0.000, 0.000, 0.707, -0.707)
	Nonuniform (u_Δ, v_Δ)	(-0.539, 0.539, 0.458, -0.458)
1948	$\Pr[G = 0]$	0.514
	$\mathscr{A}$	{0, 1, 2, 3}
	$\Pr[A \mid G = 0]$	(0.091, 0.335, 0.408, 0.166)
	$\Pr[A \mid G = 1]$	(0.037, 0.222, 0.444, 0.296)
	$\mathscr{R}$	{0, 1, 2}
	(u_0, v_0)	(0.187, -0.690, 0.503, -0.269, -0.118, 0.387)
	Uniform (u_Δ, v_Δ)	(0.000, 0.000, 0.000, -0.455, -0.360, 0.815)
	Nonuniform (u_Δ, v_Δ)	(0.207, -0.009, -0.198, 0.752, -0.188, -0.563)

4 Results

Representative simulation results for the 3 integer seeds appearing in table 3.3 are displayed in figures 4.1 through 4.4. The figures are structured as follows. The first row indicates results related to DIF detection, and the second to DIF classification. Gray lines in each plot indicate rates of MLE existence; these upper bound the performance of the asymptotic method, since the asymptotic methods issue "failure" or "unclassifiable" conclusions when MLEs fail to exist. The first row indicates MLE existence for $\mathscr{P}_{\text{no23}}$, and the second for $\mathscr{P}_{\text{no3w}}$. (Also indicated, in a lighter gray, are rates of meeting the heuristic for applying asymptotic tests which requires that at least 80% of expected counts under the MLE distribution in $\mathscr{P}_*$ be at least 5.) The colored lines in the first plot shows the proportion of tables generated using a no DIF model that were correctly classified as having no DIF (i.e., roughly one minus type I error rate for DIF detection). The colored lines in the remaining four plots of the first row indicate power for DIF detection. In the second row, colored lines in the second and third plots show the proportion of tables generated using a uniform DIF model that were correctly classified as such (i.e., roughly one minus
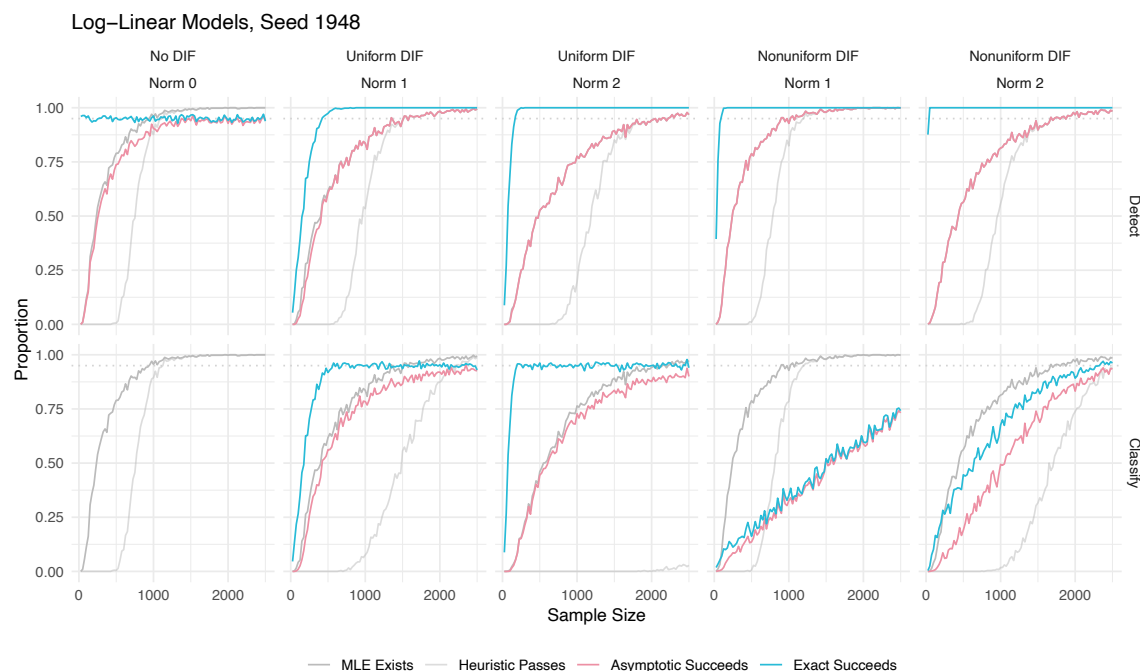
Figure 4.1

Results of DIF analysis using log-linear models on $2 \times 2 \times 2$ tables. The asymptotic and exact strategies converge rapidly in performance.

type I error for DIF classification), and in the fourth and fifth plots, they indicate power for DIF classification.

Figures 4.1 and 4.2 depict results of DIF analyses using log-linear models. Figure 4.1 shows results from a model of low dimensionality, where tables are $2 \times 2 \times 2$. MLEs frequently exist in these tables (except for the smallest of sample sizes), and the performance of asymptotic and exact methods converge quickly to each other. Type I error rates for both detection and classification converge quickly to the nominal rate of $\alpha = 0.05$. Power for DIF detection converges quickly to 100%, though power for classification takes longer. Figure 4.2 shows analogous results from a model with higher dimensionality, where tables are $4 \times 2 \times 3$. The difference between the two strategies is stark. The exact strategy holds type I error rates at nominal values and obtains substantial power even in small sample situations where MLEs are extremely unlikely to exist. In fact, in this situation MLEs frequently fail to exist even for the largest samples we generated, and the two methods do not converge in performance until the very largest sample sizes in our simulation.

Figures 4.3 and 4.4 depict results of DIF analyses using logistic regression models. They exhibit roughly the same patterns as the ones for log-linear models.

Overall, our simulations demonstrate that the exact strategies of algebraic statistics are very consistent at holding type I error rates at nominal values. Conditional p-values are known to exhibit type I error rates slightly below nominal due to discreteness of the distribution (Agresti, 2013, section 3.5.5). This conservatism is visible in our simulations,
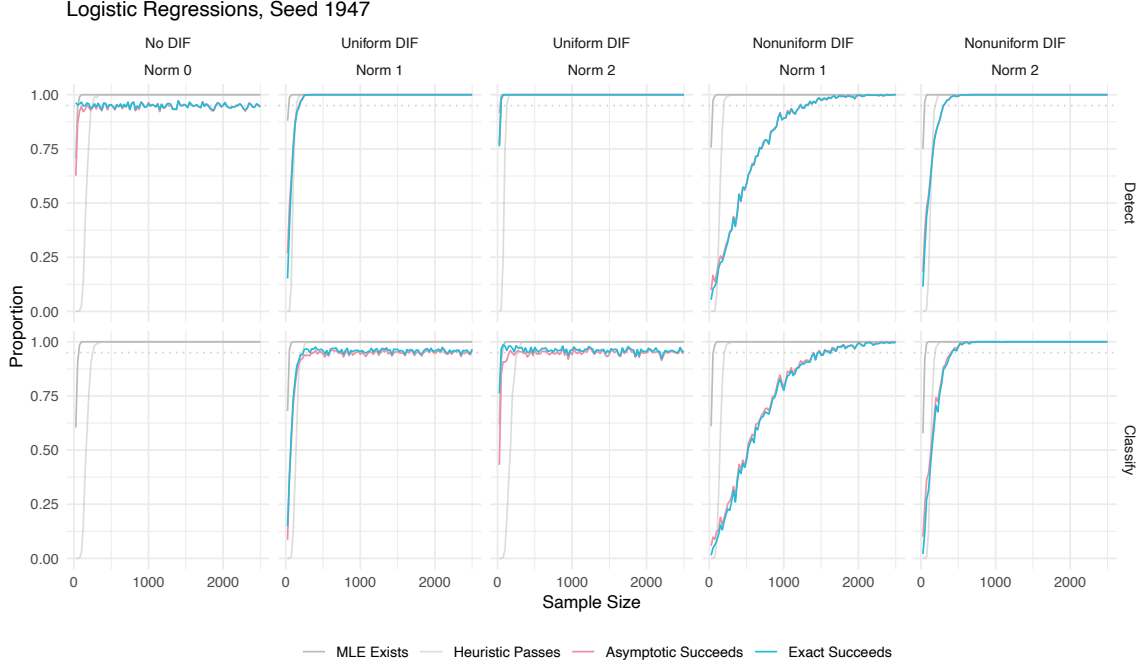
Figure 4.2

Results of DIF analysis using log-linear models on $4 \times 2 \times 3$ tables. There are sizable gaps in performance between the asymptotic and exact strategies.

but it is minimal: type I error rates for the exact methods converge rapidly to nominal. In the range of sample sizes we considered, most of the discrepancy between the exact and asymptotic strategies arises due to MLE nonexistence, which is especially prevalent for smaller sample sizes and many-celled tables, although not restricted to such cases. It is not infrequent to observe enough sampling zeros that MLEs fail to exist, even with $4 \times 2 \times 3$ tables as in figure 4.2 or $7 \times 2 \times 2$ tables as in figure 4.4, and the problem would only grow as table dimensionality increases. This bottlenecks the performance of asymptotic methods, but the exact methods of algebraic statistics are unaffected: they remain viable with competitive power and type I error control. When MLE existence is not an issue, there are typically only minor discrepancies between the two strategies.

5   Discussion

Our study has sought to demonstrate that the exact methods of algebraic statistics, and the practical implementation of these methods that is available through the R package `algstat`, may prove to be a useful tool to researchers working in psychometrics, education research, and other fields who find themselves faced with analyzing sparse multi-way contingency tables. These techniques can be used even when MLEs fail to exist, and can help circumvent the need to collapse levels in order to meet sample size requirements for applying asymptotics.

The techniques apply broadly, not only to the log-linear models and logistic re-

Figure 4.3

Results of DIF analysis using logistic regressions on $2 \times 2 \times 2$ tables. As in figure 4.1, the asymptotic and exact strategies converge rapidly in performance.

gressions we have described above. For example, there is another method of DIF analysis involving the Cochran-Mantel-Haenszel (CMH) procedure which fits squarely into the framework we have described. For this method, one takes $\mathcal{R} = \{0, 1\}$ and $\mathscr{P}$ to be the set of distributions satisfying $\operatorname{logit} \Pr[R = 1 \mid A = a, G = g] = \tau_1(a) + \tau_2 g$ for $\tau_1 \in \mathbb{R}^{\mathscr{A}}$ and $\tau_2 \in \mathbb{R}$ (cf. Swaminathan & Rogers, 1990, pp. 363–4). Then $\mathscr{P}_{\mathrm{no3w}} = \mathscr{P}$, so one cannot use this model for DIF classification, but $\mathscr{P}_{\mathrm{no23}}$ is the proper submodel of distributions where $\tau_2 = 0$, so it is still possible to use this family for DIF detection. Goodness-of-fit for the CMH procedure is typically assessed through asymptotics, but exact tests can also be used since these families are discrete exponential families of the form studied by Diaconis and Sturmfels (1998, equation (1.1)): the marginals $n_{a,g,+}$ and $n_{a,+,1}$ are sufficient statistics for $\mathscr{P}_{\mathrm{no23}}$, and this list is augmented with $n_{+,1,1}$ for $\mathscr{P}_{\mathrm{no3w}} = \mathscr{P}$.

Similarly, Bock's model (3.1) is also a discrete exponential family of the form studied by Diaconis and Sturmfels (1998, equation (1.1)).[15] Thus assessments of the fit of this model $\mathscr{P}$, and of its submodels $\mathscr{P}_{\mathrm{no23}} \subseteq \mathscr{P}_{\mathrm{no3w}}$, can also be used to detect and classify DIF. Asymptotic strategies for assessing this fit would rely on first computing an MLE, which can be done using methods described by Bock (1972), Sundberg (2019, chapter 3), or Efron (2023, section 2.6). Alternatively, one can employ the exact strategy of Diaconis and Sturmfels (1998) to assess the fit, following the pattern we have described in this study.

In fact, goodness-of-fit tests for discrete exponential families are applied in item

---

[15]In their notation, the map $T : \mathscr{A} \times \mathscr{G} \times \mathscr{R} \to (\mathbb{R}^{\mathscr{G} \times \mathscr{R}})^2 = (\mathbb{R}^2)^{\mathscr{G} \times \mathscr{R}}$ is given by $T(a, g, r) = [(g, r) \mapsto (a, 1)]$.
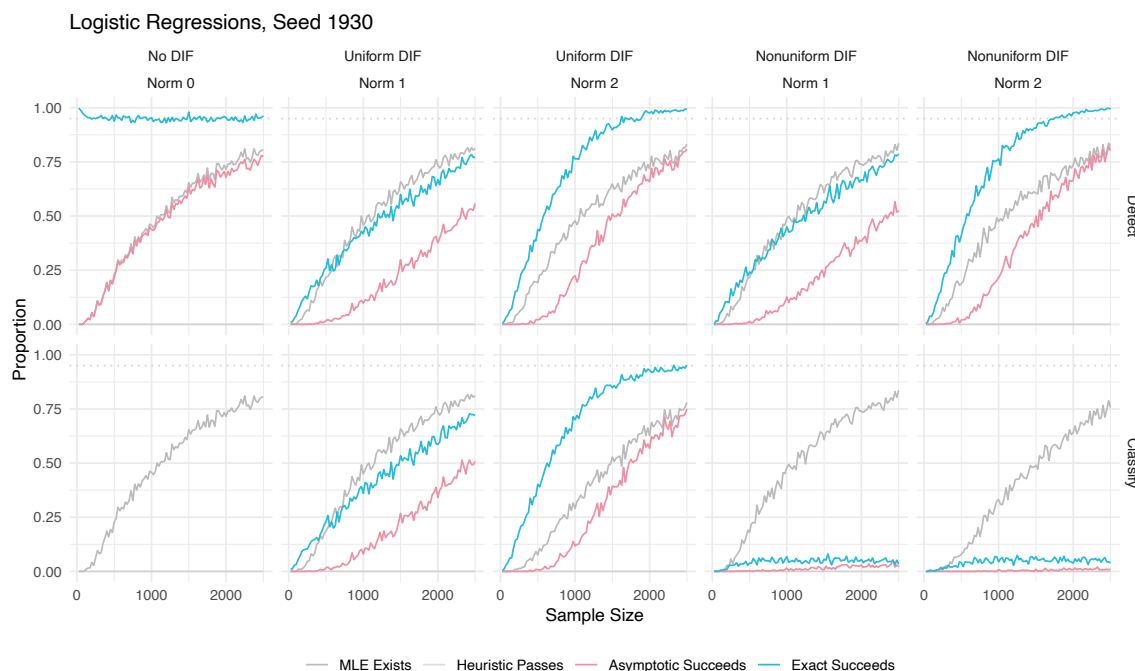
Figure 4.4

Results of DIF analysis using logistic regressions on $7 \times 2 \times 2$ tables. There are sizable gaps in performance between the asymptotic and exact strategies, as in figure 4.2. Note that classification of nonuniform DIF is extremely difficult for both the algebraic and asymptotic methods because the nonuniform DIF vector corresponding to this seed has u_Δ very close to 0 (cf. table 3.3 and the Kullback-Leibler heat map on GitHub).

response theory beyond just DIF analysis (cf. Hambleton & Swaminathan, 1985, chapters 8–9; Kelderman, 1996; Ostini & Nering, 2006, chapter 5). When data is sparse, these tests for model fit may also benefit from the techniques of algebraic statistics.

Our investigations suggest some natural avenues for future research. For example, our simulations capped the number of ability and response levels at 7, and in this range, the relevant Markov basis computations were fairly quick. Markov basis computations for the $9 \times 2 \times 2$ tables that occurred in table 2.5 were still quite feasible for a personal computer. That said, computation of Markov bases is known generally to be a bottleneck for applying the exact techniques of algebraic statistics. However, a number of techniques have been proposed in the literature to attempt to bypass this bottleneck. For example, Aoki et al. (2012, chapter 16) describes a technique in which Markov bases are replaced by lattice bases, which are easier to compute but require more care when generating the Markov chain. Alternatively, Dobra (2012) describes a technique where Markov bases are computed dynamically at each step of the Markov chain. Yet other techniques are described by Chen et al. (2006), Chen et al. (2010) and Rapallo and Yoshida (2010). It would be interesting to investigate applications of these techniques to DIF analysis and item response theory, to other areas in psychometrics and education research, and to the sciences more broadly.

References

4ti2 Team. (2013). *4ti2—A software package for algebraic, geometric and combinatorial problems on linear spaces*. http://www.4ti2.de

Agresti, A. (2013). *Categorical data analysis* (Third). John Wiley & Sons.

Albert, A., & Anderson, J. A. (1984). On the existence of maximum likelihood estimates in logistic regression models. *Biometrika*, *71*, 1–10. https://doi.org/10.1093/biomet/71.1.1

Aoki, S., Hara, H., & Takemura, A. (2012). *Markov bases in algebraic statistics*. New York, NY: Springer. https://doi.org/10.1007/978-1-4614-3719-2

Baldoni, V., Berline, N., De Loera, J., Dutra, B., Koppe, M., Moreinis, S., Pinto, G., Vergne, M., & Wu, J. (2013). *A user's guide for LattE integrale v1.7.2*. http://www.math.ucdavis.edu/~latte/

Barndorff-Nielsen, O. (2014). *Information and exponential families in statistical theory* [Reprint of the 1978 original]. John Wiley & Sons, Ltd., Chichester. https://doi.org/10.1002/9781118857281

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, *57*(1), 289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika*, *37*(1), 29–51. https://doi.org/10.1007/BF02291411

Bogdan, K., Bosy, M., & Skalski, T. (2022). Maximum likelihood estimation for discrete exponential families and random graphs. *ALEA, Lat. Am. J. Probab. Math. Stat.*, *19*(1), 1045–1070. https://doi.org/10.30757/ALEA.v19-43

Boyett, J. M. (1979). Algorithm AS 144: Random $R \times C$ tables with given row and column totals. *Applied Statistics*, *28*(3), 329. https://doi.org/10.2307/2347219

Casella, G., & Berger, R. L. (2001). *Statistical inference* (Second). Duxbury-Thomson Learning.

Chen, Y., Dinwoodie, I. H., & Sullivant, S. (2006). Sequential importance sampling for multiway tables. *The Annals of Statistics*, *34*(1). https://doi.org/10.1214/009053605000000822

Chen, Y., Dinwoodie, I. H., & Yoshida, R. (2010). Markov chains, quotient ideals and connectivity with positive margins. In *Algebraic and geometric methods in statistics* (pp. 99–110). Cambridge Univ. Press, Cambridge.

Clarkson, D. B., Fan, Y.-a., & Joe, H. (1993). A remark on algorithm 643: FEXACT. *ACM Transactions on Mathematical Software*, *19*(4), 484–488. https://doi.org/10.1145/168173.168412

Cox, D. A., Little, J., & O'Shea, D. (2015). *Ideals, varieties, and algorithms. An introduction to computational algebraic geometry and commutative algebra* (4th revised ed.). Cham: Springer. https://doi.org/10.1007/978-3-319-16721-3

Dancer, L. S., Anderson, A. J., & Derlin, R. L. (1994). Use of log-linear models for assessing differential item functioning in a measure of psychological functioning. *Journal of*

Consulting and Clinical Psychology, *62*(4), 710–717. https://doi.org/10.1037/0022-006x.62.4.710

Deckert, A. C., & Kummerfeld, E. (2019). Investigating the effect of binning on causal discovery. *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. https://doi.org/10.1109/bibm47256.2019.8983336

Diaconis, P., & Sturmfels, B. (1998). Algebraic algorithms for sampling from conditional distributions. *The Annals of Statistics*, *26*(1). https://doi.org/10.1214/aos/1030563990

Dobra, A. (2012). Dynamic Markov bases. *J. Comput. Graph. Statist.*, *21*(2), 496–517. https://doi.org/10.1080/10618600.2012.663285

Drton, M., Sturmfels, B., & Sullivant, S. (2009). *Lectures on algebraic statistics* (Vol. 39). Basel: Birkhäuser.

Efron, B. (2023). *Exponential families in theory and practice* (Vol. 16). Cambridge: Cambridge University Press. https://doi.org/10.1017/9781108773157

Eriksson, N., Fienberg, S. E., Rinaldo, A., & Sullivant, S. (2006). Polyhedral conditions for the nonexistence of the MLE for hierarchical log-linear models. *Journal of Symbolic Computation*, *41*(2), 222–233. https://doi.org/10.1016/j.jsc.2005.04.003

Fienberg, S. E., & Rinaldo, A. (2007). Three centuries of categorical data analysis: Log-linear models and maximum likelihood estimation. *Journal of Statistical Planning and Inference*, *137*(11), 3430–3445. https://doi.org/10.1016/j.jspi.2007.03.022

Fisher, R. A. (1922). On the interpretation of χ^2 from contingency tables, and the calculation of *p*. *Journal of the Royal Statistical Society*, *85*(1), 87–94. https://doi.org/https://doi.org/10.1098/rsta.1922.0009

Fisher, R. A., & Russell, E. J. (1922). On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, *222*(594–604), 309–368. https://doi.org/10.1098/rsta.1922.0009

French, A. W., & Miller, T. R. (1996). Logistic regression and its use in detecting differential item functioning in polytomous items. *Journal of Educational Measurement*, *33*(3), 315–332. https://doi.org/10.1111/j.1745-3984.1996.tb00495.x

Haberman, S. J. (1972). Algorithm AS 51: Log-linear fit for contingency tables. *Applied Statistics*, *21*(2), 218. https://doi.org/10.2307/2346506

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory*. Springer Netherlands. https://doi.org/10.1007/978-94-017-1988-9

Hastings, W. K. (1970). Monte carlo sampling methods using markov chains and their applications. *Biometrika*, *57*(1), 97–109. https://doi.org/10.1093/biomet/57.1.97

Jacobson, M. (1989). Existence and unicity of MLEs in discrete exponential family distributions. *Scand. J. Stat.*, *16*(4), 335–349.

Kahle, D., Garcia-Puente, L., & Yoshida, R. (2017a). `algstat`: *Algebraic statistics in R* [R package version 0.1.1]. https://github.com/dkahle/algstat

Kahle, D., Garcia-Puente, L., & Yoshida, R. (2017b). *Latte: `LattE` and `4ti2` in R* [R package version 0.2.0]. https://github.com/dkahle/latte

Kahle, D., Yoshida, R., & Garcia-Puente, L. (2018). Hybrid schemes for exact conditional inference in discrete exponential families. *Annals of the Institute of Statistical Mathematics*, *70*(5), 983–1011. https://doi.org/10.1007/s10463-017-0615-z

Kelderman, H. (1996). Multidimensional rasch models for partial-credit scoring. *Applied Psychological Measurement*, *20*(2), 155–168. https : / / doi . org / 10 . 1177 / 014662169602000205

Konis, K. (2007). *Linear programming algorithms for detecting separated data in binary logistic regression models* [Doctoral dissertation, University of Oxford]. https://ora. ox.ac.uk/objects/uuid:8f9ee0d0-d78e-4101-9ab4-f9cbceed2a2a

Lauritzen, S. L. (1996). *Graphical models.* (Vol. 17). Oxford: Oxford Univ. Press.

Martinkova, P., & Drabinova, A. (2018). ShinyItemAnalysis for teaching psychometrics and to enforce routine analysis of educational tests. *The R Journal*, *10*(2), 503–515. https://doi.org/10.32614/RJ-2018-074

McFarland, J. L., Price, R. M., Wenderoth, M. P., Martinkova, P., Cliff, W., Michael, J., Modell, H., & Wright, A. (2017). Development and Validation of the Homeostasis Concept Inventory. *CBE Life Sciences Education*, *16*(2), ar35. https://doi.org/10. 1187/cbe.16-10-0305

Mehta, C. R., & Patel, N. R. (1986). Algorithm 643. *ACM Transactions on Mathematical Software*, *12*(2), 154–161. https://doi.org/10.1145/6497.214326

Mellenbergh, G. J. (1982). Contingency table models for assessing item bias. *Journal of Educational Statistics*, *7*(2), 105. https://doi.org/10.2307/1164960

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., & Teller, E. (1953). Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, *21*(6), 1087–1092. https://doi.org/10.1063/1.1699114

Ostini, R., & Nering, M. (2006). *Polytomous item response theory models*. SAGE Publications, Inc. https://doi.org/10.4135/9781412985413

Patefield, W. M. (1981). Algorithm AS 159: An efficient method of generating random $R \times C$ tables with given row and column totals. *Applied Statistics*, *30*(1), 91. https: //doi.org/10.2307/2346669

Pearson, K. (1900). On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Philosophical Magazine Series 5*, *50*(302), 157–175. https://doi.org/10.1080/14786440009463897

R Core Team. (2023). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria. https://www.R-project.org/

Rapallo, F., & Yoshida, R. (2010). Markov bases and subbases for bounded contingency tables. *Ann. Inst. Stat. Math.*, *62*(4), 785–805. https://doi.org/10.1007/s10463-010-0289-2

Samejima, F. (1968). Estimation of latent ability using a response pattern of graded scores. *ETS Research Bulletin Series*, *1968*(1), i–169. https://doi.org/https://doi.org/10. 1002/j.2333-8504.1968.tb00153.x

Scheuneman, J. (1979). A method of assessing bias in test items. *Journal of Educational Measurement*, *16*(3), 143–152. https://www.jstor.org/stable/1433816

Simpson, E. H. (1951). The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, *13*(2), 238–241. https: //doi.org/10.1111/j.2517-6161.1951.tb00088.x

Sturmfels, B. (1996). *Gröbner bases and convex polytopes* (Vol. 8). Providece, RI: AMS, American Mathematical Society.

Sundberg, R. (2019). *Statistical modelling by exponential families* (Vol. 12). Cambridge: Cambridge University Press. https://doi.org/10.1017/9781108604574

Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement*, *27*(4), 361–370. http://www.jstor.org/stable/1434855

van der Flier, H., Mellenbergh, G. J., Adèr, H. J., & Wijn, M. (1984). An iterative item bias detection method. *Journal of Educational Measurement*, *21*(2), 131–145. https://doi.org/10.1111/j.1745-3984.1984.tb00225.x

van de Vijver, F. J. R., & Leung, K. (2021, June). *Methods and data analysis for cross-cultural research* (V. H. Fetvadjiev, J. R. J. Fontaine, & J. He, Eds.). Cambridge University Press. https://doi.org/10.1017/9781107415188

Wilks, S. S. (1938). The large-sample distribution of the likelihood ratio for testing composite hypotheses. *The Annals of Mathematical Statistics*, *9*(1), 60–62. https://doi.org/10.1214/aoms/1177732360

Yee, D., & Ho, A. (2015). Discreteness causes bias in percentage-based comparisons: A case study from educational testing. *The American Statistician*, *69*(3), 174–181. https://doi.org/10.1080/00031305.2015.1031828

Yesiltas, G., & Paek, I. (2020). A log-linear modeling approach for differential item functioning detection in polytomously scored items. *Educational and Psychological Measurement*, *80*(1), 145–162. https://doi.org/10.1177/0013164419853000

Yule, G. U. (1903). Notes on the theory of association of attributes in statistics. *Biometrika*, *2*(2), 121–134. https://doi.org/10.1093/biomet/2.2.121