


On the Construction of Unbiased Estimators for the Group Testing Problem

Gregory Haber^{} and Yaakov Malinovsky

University of Maryland, Baltimore County, Baltimore, USA

Abstract

Debiased estimation has long been an area of research in the group testing literature. This has led to the development of several estimators with the goal of bias minimization and, recently, an unbiased estimator based on sequential binomial sampling. Previous research, however, has focused heavily on the simple case where no misclassification is assumed and only one trait is to be tested. In this paper, we consider the problem of unbiased estimation in these broader areas, giving constructions of such estimators for several cases. We show that, outside of the standard case addressed previously in the literature, it is impossible to find any proper unbiased estimator, that is, an estimator giving only values in the parameter space. This is shown to hold generally under any binomial or multinomial sampling plans.

AMS (2000) subject classification. Primary 62F10; Secondary 62L12.

Keywords and phrases. Binomial sampling plans, Group testing, Multinomial sampling plans, Sequential estimation, Unbiased estimation

1 Introduction

Group testing, which includes generally any situation in which specimens are tested in groups instead of individually, has been an ongoing area of research in the statistical literature for over 70 years. First introduced in Dorfman (1943) as a means of screening U.S. Army inductees for syphilis, subsequent research has led to the development of two overarching fields, case identification (as in Dorfman's original work) and estimation.

The estimation problem, which is the focus of the current work, has as its prototypical case the prevalence estimation of a single binary trait from an assumed infinite population when testing is done error free. Typically the trait of interest will be rare, so that grouping can lead to a significant reduction in the number of tests required or an increase in efficiency (in terms of mean square error) for a fixed number of trials.

While the above scenario is important theoretically, in many applications tests will be subject to misclassification error which must be accounted for when analyzing group testing data. Research in this area has been broad, covering an array of cases (see, as a few examples, Tu et al. (1995), Hung and Swallow (1999), Liu et al. (2012), McMahan et al. (2013), Zhang et al. (2014), Huang et al. (2017), & Li et al. (2017)).

An additional area which is becoming increasingly important for applications due to the growing development of multiplex screening tools is the estimation of prevalences for several traits simultaneously. Such assays can be modelled naturally using multinomial sampling and extensions of group testing methods to such designs can be found in Hughes-Oliver and Rosenberger (2000), Pfeiffer et al. (2002), Tebbs et al. (2013), Ding and Xiong (2015), & Warasi et al. (2016).

In all cases, one of the major difficulties in carrying out estimation using group testing is that the standard maximum likelihood estimator (MLE) is biased, often quite significantly depending on the true underlying prevalence and group sizes (see, for example, Gibbs and Gower (1960), Thompson (1962), & Swallow (1985)). This has led to the development of several alternative debiased estimators, that is, estimators with significantly reduced bias relative to the MLE (see Burrows (1987), Tebbs et al. (2003), Hepworth and Watson (2009), Ding and Xiong (2016), & Santos and Dorgman (2016)). Perhaps most importantly, such estimators generally yield large reductions in the mean square error (MSE) when compared with the MLE, indicating the importance of developing such tools for group testing data. While the bias and MSE can be controlled to a degree with good design (i.e. appropriate choices of the group sizes), this usually requires the use of prior knowledge regarding the prevalence parameter or adaptive designs which may not be feasible in many cases (see, for example, Chiang and Reeves (1962), Hughes-Oliver and Swallow (1994), & Haber and Malinovsky (2017)).

It should be noted that, if fixed binomial sampling is used, it is impossible to find any unbiased estimator for the underlying parameter when group testing is used. This fact was mentioned in Hall (1963) and follows from a general result concerning the estimation of the function of a binomial parameter given in Lehmann and Casella (1998), p. 100. This result can be easily extended to the cases when misclassification is present and/or multiple traits are screened simultaneously, so that an unbiased estimator cannot exist in such cases when fixed sampling (either binomial or multinomial, as appropriate) is used.

To consider unbiased estimation for the prevalence in the group testing problem, then, it is necessary to consider the broader class of binomial and multinomial sampling plans (defined in the following section), of which the fixed binomial and multinomial designs are members. In a recent work, Haber et al. (2018) took this approach and showed, based on results from DeGroot (1959), that under a certain class of inverse binomial sampling models it is possible to construct an unbiased estimator. Their work, however, was restricted to the simple case outlined above where only a single trait is to be estimated without misclassification.

In this paper, we extend the question of unbiased estimation for group testing to the above generalizations, misclassification and multiple-trait screening. In particular, we focus on the case when misclassification errors are assumed known and on the simultaneous estimation of two correlated diseases.

We show that, in both cases, unbiased estimation is possible using inverse sampling and constructions are provided under the appropriate models. It is shown, however, that these estimators are improper, that is, they lie outside of the parameter space for some sample values. The core theoretical result of this work is to show that this will be true for any unbiased estimator under any binomial sampling plan with misclassification or any multinomial sampling plan (with at least three elements), even with perfect testing.

2 Binomial and Multinomial Sampling Plans

In this section, we define the general classes of binomial and multinomial (of which binomial is a special case) sampling plans. A more detailed treatment of binomial sampling plans can be found in, among others, Girshick et al. (1946) & DeGroot (1959). Similar results for multinomial sampling plans can be found in Kremers (1990) & Koike (1993).

In general, a binomial sampling plan \mathcal{S} is a set of points on the non-negative xy -plane determined by a set of boundary points $\mathcal{B}_{\mathcal{S}}$. For all plans, sampling begins at the origin and increases the x or y coordinate with probabilities θ and $1 - \theta$, respectively, iteratively until a point in $\mathcal{B}_{\mathcal{S}}$ is reached. This class is very broad, and includes both the fixed binomial and inverse binomial sampling plans, as well as many variations of bounded or fully sequential sampling designs.

The class of multinomial sampling plans is a direct generalization of the above idea. We say \mathcal{S}_t is a multinomial sampling plan in $t + 1$ dimensions if \mathcal{S}_t is a set of points on the non-negative orthant lying in $t + 1$ dimensional space. The plan is similarly determined by a set of boundary points $\mathcal{B}_{\mathcal{S}_t}$. Sampling begins at the origin and increases the x_i coordinate, $i = 0, 1, \dots, t$,

at each step with probability $\theta_i, i = 0, 1, \dots, t$ where $\boldsymbol{\theta} = (\theta_1, \theta_2, \dots, \theta_t)$ is a multinomial parameter and $\theta_0 = 1 - \mathbf{1}'\boldsymbol{\theta}$.

3 Unbiased Estimation Under Inverse Multinomial Sampling

In this section a theorem with necessary and sufficient conditions for unbiased estimation of a function of the parameter vector of an inverse multinomial model is given. This is a generalization of Theorem 4.1 found in DeGroot (1959), which applies only for two class problems, and will be used in subsequent sections to construct unbiased estimators under group testing models for one and two traits. While the results presented here are applicable in many situations, for convenience we refer to testing for single or multiple diseases throughout.

Let $\boldsymbol{\mu} = (\mu_1, \dots, \mu_t)'$ with $\mu_0 = 1 - \mathbf{1}'\boldsymbol{\mu}$ and let $IMN_t(c, \boldsymbol{\mu})$ denote the t -class inverse multinomial model with parameter $\boldsymbol{\mu}$ which samples until c observations from the class corresponding to μ_0 are observed. Then, the random variable $\mathbf{X} \sim IMN_t(c, \boldsymbol{\mu})$ with parameter space $\bar{\Psi} = \{\boldsymbol{\mu} : \mathbf{1}'\boldsymbol{\mu} < 1, 0 \prec \boldsymbol{\mu} \prec 1\}$ where \prec denotes element-wise inequality, if

$$P(\mathbf{X} = \mathbf{x}) = \binom{c + \sum_{i=1}^t x_i - 1}{c - 1, x_1, \dots, x_t} \mu_0^c \prod_{i=1}^t \mu_i^{x_i}.$$

Let $\Psi \subset \bar{\Psi}$ with $int(\Psi)$ the interior of Ψ . Then, we say that X has an inverse multinomial distribution with restricted parameter space if X has the same pdf as above but with parameter space Ψ . A special case of this distribution with $t = 1$ is the inverse binomial, which corresponds to the classical group testing problem when screening for one disease.

Theorem 1. *Let $\mathbf{X} \sim IMN_t(c, \boldsymbol{\mu})$ with restricted parameter space Ψ . A function $h(\boldsymbol{\mu})$ is estimable unbiasedly for all $\boldsymbol{\mu} \in int(\Psi)$ if and only if h is an analytic function on a region containing an open-ball about $\mathbf{0} \in \mathbb{R}^t$ and $int(\Psi)$. The estimator is given by*

$$f(\mathbf{x}) = \frac{(c - 1)!}{(c + \sum_{i=1}^t x_i - 1)!} \left. \frac{\partial^{\sum_{i=1}^t x_i} g(\boldsymbol{\mu})}{\partial \mu_1^{x_1} \dots \partial \mu_t^{x_t}} \right|_{\boldsymbol{\mu}=\mathbf{0}},$$

where $g(\boldsymbol{\mu}) = \frac{h(\boldsymbol{\mu})}{\mu_0^c}$.

Remark 1. *It should be noted that the restriction to values in $int(\Psi)$ is not absolute, and values from the boundary such as $\mu_0 = 1$ may be estimable unbiasedly, while others such as values in the plane $\mu_0 = 0$ can not. The*

latter point can be seen by noting that the function $g(\boldsymbol{\mu})$ above is undefined when $\mu_0 = 0$. In general, there is little interest in estimating any points on the boundary, hence the restriction to the interior is sufficient.

While we will be interested here in applying this theorem only in the context of group testing, other possibilities include the unbiased estimation of the relative risk or odds ratio of two diseases estimated simultaneously.

4 One Disease Case with Misclassification

For the single disease group testing problem, we assume an infinite population of individuals whose binary status can be represented by independent random variables $\varphi \sim Ber(p)$. In what follows, p is the quantity we seek to estimate. If, instead of as individuals, members of this population are tested in groups of size k , we have the new random variable $\vartheta = \max\{\varphi_1, \dots, \varphi_k\} \sim Ber(1 - q^k)$, where $q = 1 - p$.

To incorporate testing error, let $\tilde{\vartheta}$ be the true, latent value of the observed ϑ . Then, we define the specificity and sensitivity, respectively, as $\pi_0 = P(\vartheta = 0 | \tilde{\vartheta} = 0)$ and $\pi_1 = P(\vartheta = 1 | \tilde{\vartheta} = 1)$. This yields the distribution $\vartheta \sim Ber(\theta)$ where $\theta = \pi_1 - \nu q^k$ with $\nu = \pi_1 + \pi_0 - 1$. It should be noted that this model is identifiable if and only if $\nu \neq 0$. A standard assumption to address this, which is made here as well, is that both π_0 and π_1 are greater than 0.5. This is reasonable as it merely assumes the test performs better than random guessing.

To find an unbiased estimator using Theorem 1, we consider $Y \sim IMN_1(c, \theta)$ which is the number of positives until c groups testing negative for the disease are observed. Note that the parameter space here is restricted as a function of θ when either $\pi_0 < 1$ or $\pi_1 < 1$ since, for $0 < p < 1$, $1 - \pi_0 < \theta < \pi_1$. Then, we seek an unbiased estimator of $q = h(\theta) = \frac{(\pi_1 - \theta)^{1/k}}{\nu^{1/k}}$, which is analytic on the interval $|\theta| < \pi_1$.

Result 1. *For the one disease case with misclassification, with $Y \sim IMN_1(c, \pi_1 - \nu q^k)$, the unique unbiased estimator of p is given by*

$$\hat{p}_{UB}(y) = 1 - \left(\frac{\pi_1}{\nu}\right)^{1/k} \sum_{i=0}^y \binom{y}{i} \frac{1}{\pi_1^{y-i}} \frac{(c+i-1)!}{(c+y-1)!} \prod_{j=i}^y \frac{(y-j-1/k)}{(y-i-1/k)}, y=0, 1, 2, \dots$$

This result can be used to derive an unbiased estimator for the perfect testing case ($\pi_0 = \pi_1 = 1$) as given in the following corollary and Haber et al. (2018).

Corollary 1. *For the one disease case with no misclassification, an unbiased estimator of p is given by*

$$\hat{p}_{UB}(y) = 1 - \frac{1}{\left(1 - \frac{1}{k(c+y)}\right)} \prod_{i=0}^y \left(1 - \frac{1}{k(c+i)}\right), y = 0, 1, 2, \dots$$

4.1. Non-Properness of Unbiased Estimator While the estimator given in Result 1 is unbiased, for either $\pi_0 < 1$ or $\pi_1 < 1$ it is an improper estimator, that is it yields values lying outside the parameter space.

To see this, note that if $\pi_0 < 1$ then, for any π_1 , we have $\hat{p}_{UB}(0) = 1 - \left(\frac{\pi_1}{\pi_1 + \pi_0 - 1}\right)^{1/k} < 0$. Likewise, If $\pi_0 = 1$, then, for $\pi_1 < 1$ and $y \geq 1$,

$$\hat{p}_{UB}(y) = \frac{1}{k} \sum_{i=0}^{y-1} \binom{y}{i} \frac{1}{\pi_1^{y-i}} \frac{(c+i-1)!}{(c+y-1)!} \prod_{j=i}^{y-1} \frac{(y-j-1/k)}{(y-i-1/k)}.$$

Now, each term of the sum in this expression is positive, so it is sufficient to show that for some y at least one term is greater than 1, resulting in a total estimate larger than 1. For the $i = 0$ term, we have $\frac{1}{kc} \prod_{j=1}^{y-1} \frac{1}{\pi_1} \left(1 - \frac{c+1/k}{c+y-j}\right)$

which diverges since $\pi_1 < 1$.

While these results, combined with the necessity clause of Theorem 1, mean that there exist no proper unbiased estimators under the inverse binomial model, in the following result we extend this idea to show that no such estimator exists under any binomial sampling plan when misclassification is present.

Theorem 2. *Let \mathcal{S} be a binomial sampling plan with set of boundary points $\mathcal{B}_{\mathcal{S}}$ for which, at a given step, the x and y coordinates are increased with probability $\theta = \pi_1 - \nu q^k$ and $1 - \theta$, respectively. Then, if $\pi_0 < 1$ or $\pi_1 < 1$, there exists no proper unbiased estimator of p under \mathcal{S} .*

It should be noted that, while the explicit construction of the unbiased estimator in Result 1 required the assumption that the misclassification parameters were known, the result of Theorem 2 holds more generally, even when this assumption does not hold.

From the proof of Theorem 2 we get the following corollary, which is also given in Haber et al. (2018), showing that the above inverse binomial model

which counts until c negatives is the only one yielding an unbiased estimator of p .

Corollary 2. *Let $Y \sim IMN_1(c, 1 - \theta)$ where $\theta = \pi_1 - \nu q^k$, so that Y is the number of negative groups drawn until c positive results are observed. Then, there exists no unbiased estimator of p for any values of π_0 and π_1 .*

We end this section with a numerical illustration of the impact of the improperness property for the specific estimator given in Result 1. Figure 1 shows the relative bias, defined as $\frac{E[\hat{p}-p]}{p}$ for an estimator \hat{p} , for the unbiased estimator in Result 1, the same estimator truncated to be in the parameter space (so that it is a proper estimator), and the MLE for the model considered at the beginning of the section. The plot is given in two parts, dividing the range of p into small and medium/large due to the step increase in bias for some estimators as p nears zero. The large difference in scale should be noted when interpreting the plot.

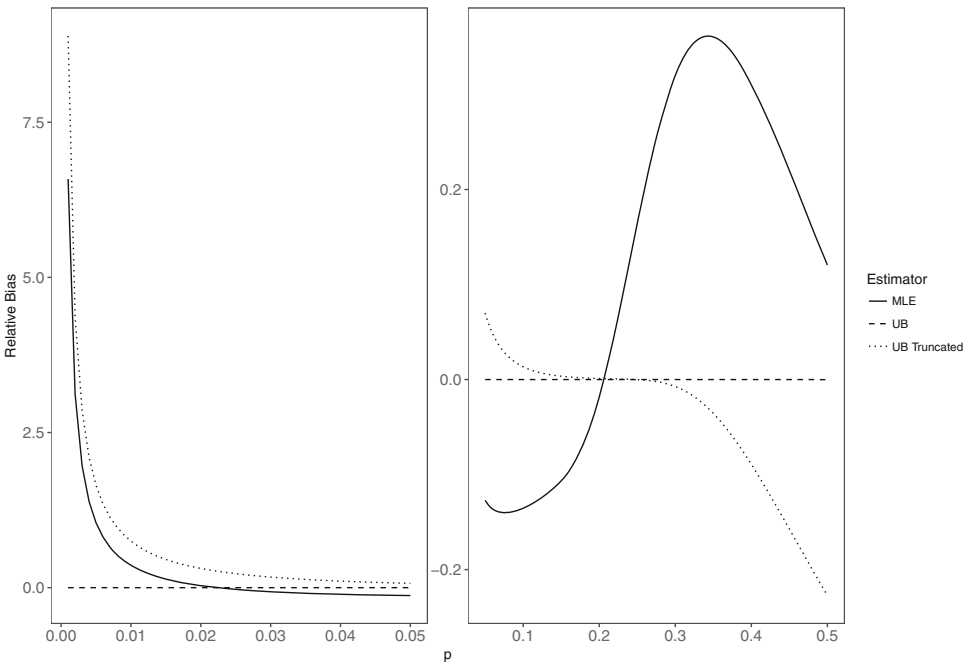


Figure 1: Relative bias, $\frac{E[\hat{p}-p]}{p}$, for the maximum likelihood estimator (MLE), unbiased estimator (UB), and the unbiased estimator truncated to be in the parameter space (UB Truncated). Here, $c = 5$, $k = 10$, $\pi_0 = 0.9$, $\pi_1 = 0.95$. Note the difference in scale for both parts of the figure

From the figure, we see that truncating the unbiased estimator has caused the bias to be increased away from zero, in some cases with a larger magnitude than that of the MLE. This shows, at least for the specific scenario considered here, that the problem of improperness is non-trivial.

5 Two Disease Case with no Misclassification

For the case of two diseases, let φ_1 and φ_2 be marginally binomial random variables with parameters $0 < p_1 < 1$ and $0 < p_2 < 1$ respectively. Then, (φ_1, φ_2) has a one-to-one correspondence to the vector $\boldsymbol{\varphi} = (\varphi_{00}, \varphi_{10}, \varphi_{01}, \varphi_{11})$ with joint multinomial distribution $\boldsymbol{\varphi} \sim MN_3(1, \mathbf{p})$ and sample space $\boldsymbol{\Psi}_{\mathbf{p}} = \{\mathbf{p} : \mathbf{1}'\mathbf{p} < 1, 0 \prec \mathbf{p} \prec 1\}$, where $\mathbf{p} = (p_{10}, p_{01}, p_{11})$ and $p_{00} = 1 - \mathbf{1}'\mathbf{p}$. Note that the marginal parameters can be expressed as $p_1 = p_{10} + p_{11}$ and $p_2 = p_{01} + p_{11}$.

If we assume no misclassification, we have the i th grouped sample $(\vartheta_{1i}^{(k)}, \vartheta_{2i}^{(k)}) = (\max\{\varphi_{1i_1}, \dots, \varphi_{1i_k}\}, \max\{\varphi_{2i_1}, \dots, \varphi_{2i_k}\})$ which corresponds to

$$\boldsymbol{\vartheta}_i^{(k)} = (\vartheta_{00}, \vartheta_{10}, \vartheta_{01}, \vartheta_{11}) \sim MN_3(1, \boldsymbol{\theta}),$$

where

$$\begin{aligned} \boldsymbol{\theta} &= (\theta_{10}, \theta_{01}, \theta_{11}) \\ &= ((p_{00} + p_{10})^k - p_{00}^k, (p_{00} + p_{01})^k - p_{00}^k, 1 - (p_{00} + p_{10})^k \\ &\quad - (p_{00} + p_{01})^k + p_{00}^k) \end{aligned} \tag{5.1}$$

and

$$\theta_{00} = 1 - \mathbf{1}'\boldsymbol{\theta} = p_{00}^k. \tag{5.2}$$

If we sample until c groups are found without either disease, and set $\mathbf{Z} = (z_{10}, z_{01}, z_{11})$ to be the sum of the observed $\boldsymbol{\vartheta}_i^{(k)}$ s, we have $\mathbf{Z} \sim IMN_3(c, \boldsymbol{\theta})$. Note that the parameter space of \mathbf{Z} , $\boldsymbol{\Psi}_{\mathbf{Z}} = \{\boldsymbol{\theta}(\mathbf{p}) : \mathbf{1}'\mathbf{p} < 1, 0 \prec \mathbf{p} \prec 1\}$, is a proper subset of the full parameter space $\boldsymbol{\Psi}_{\boldsymbol{\theta}} = \{\boldsymbol{\theta} : \mathbf{1}'\boldsymbol{\theta} < 1, 0 \prec \boldsymbol{\theta} \prec 1\}$. This fact will play a crucial role below in showing that there exists no proper unbiased estimator of \mathbf{p} .

To find an unbiased estimator, the following lemma will be needed, which is simply the result of inverting (5.1) and (5.2).

Lemma 1. *The unique function $h : \boldsymbol{\theta} \mapsto \mathbf{p}$ is given by*

$$\begin{aligned} p_{00} &= h_{00}(\boldsymbol{\theta}) = (1 - \theta_{10} - \theta_{01} - \theta_{11})^{1/k}, \\ p_{10} &= h_{10}(\boldsymbol{\theta}) = (1 - \theta_{01} - \theta_{11})^{1/k} - h_{00}(\boldsymbol{\theta}), \\ p_{01} &= h_{01}(\boldsymbol{\theta}) = (1 - \theta_{10} - \theta_{11})^{1/k} - h_{00}(\boldsymbol{\theta}), \\ p_{11} &= h_{11}(\boldsymbol{\theta}) = 1 - p_{00} - p_{10} - p_{01}. \end{aligned}$$

The function $h(\boldsymbol{\theta})$ given in Lemma 1 is analytic on an open region containing $\mathbf{0} \cup \text{int}(\Psi_Z)$, so the conditions of Theorem 1 hold and an unbiased estimator exists.

Result 2. *The unique unbiased estimator of \mathbf{p} where $\mathbf{Z} \sim \text{IMN}_3(c, \boldsymbol{\theta})$, with $\boldsymbol{\theta}$ as in Eq. 5.1, is given by*

$$\begin{aligned} \hat{p}_{00} &= \frac{1}{\left(1 - \frac{1}{k(c+z_{10}+z_{01}+z_{11})}\right)} \prod_{j=0}^{z_{10}+z_{01}+z_{11}} \left(1 - \frac{1}{k(c+j)}\right), \\ \hat{p}_{10} &= \frac{1}{\left(1 - \frac{1}{k(c+z_{10}+z_{01}+z_{11})}\right)} \prod_{j=0}^{z_{01}+z_{11}} \left(1 - \frac{1}{k(c+z_{10}+j)}\right) - \hat{p}_{00}, \\ \hat{p}_{01} &= \frac{1}{\left(1 - \frac{1}{k(c+z_{10}+z_{01}+z_{11})}\right)} \prod_{j=0}^{z_{10}+z_{11}} \left(1 - \frac{1}{k(c+z_{01}+j)}\right) - \hat{p}_{00}, \\ \hat{p}_{11} &= 1 - \hat{p}_{00} - \hat{p}_{10} - \hat{p}_{01}. \end{aligned}$$

The unbiased estimator given in Result 2 is an improper estimator. This can be shown by counterexample, considering the point $\mathbf{z} = (1, 1, 0)$. We have, evaluating at this point,

$$\begin{aligned} \hat{p}_{00} + \hat{p}_{10} + \hat{p}_{01} &= 2 \left(1 - \frac{1}{k(c+1)}\right) - \left(1 - \frac{1}{kc}\right) \left(1 - \frac{1}{k(c+1)}\right) \\ &= 1 + \frac{1}{kc} - \frac{1}{k(c+1)} - \frac{1}{k^2c(c+1)} \\ &= 1 + \frac{1}{kc} \left(1 - \frac{(c+1/k)}{c+1}\right) \\ &> 1, \end{aligned}$$

for any c and $k > 1$.

As in Theorem 2, this property can be shown to hold for any unbiased estimator under any multinomial sampling plan.

Theorem 3. *Let \mathcal{S}_3 be a multinomial sampling plan in four dimensions with boundary points $\mathcal{B}_{\mathcal{S}_3}$ such that at each step the i th coordinate, $i = 0, 1, 2, 3$, is increased with probability θ_i as in Eqs. 5.1 and 5.2. Then, there exists no proper unbiased estimator of \mathbf{p} under \mathcal{S}_3 .*

6 Two Disease Case with Misclassification

In this section we consider the two disease testing problem when misclassification is present, by looking at two models for incorporating such testing

errors. In both cases we will assume the misclassification parameters to be known a priori, although the results on non-proper estimators will hold more generally.

The first model, as introduced in Li et al. (2017), is very general, requiring no assumptions on how the marginal testing errors are combined. The downside, as we shall see, is that this requires a large number of parameters, knowledge of which may not be available for many of the assays used in applications. Let $\tilde{\varphi}_a$, $a \in \{00, 10, 01, 11\}$ be the true latent value of the observed random variable φ_a . Then, we have the misclassification parameters $\pi_{a|b} = P(\varphi_a|\tilde{\varphi}_b)$, $a, b \in \{00, 10, 01, 11\}$. While this indicates 16 parameters, each one can be expressed as a linear combination of three others, so that the model consists of twelve extra parameters.

If we again let \mathbf{Z} be the sum of the $\vartheta_i^{(k)}$ s until c groups are observed without either disease then, with the above misclassification values, and $\boldsymbol{\theta}$ as in Eqs. 5.1 and 5.2, we now have $\mathbf{Z} \sim IMN_3(c, \boldsymbol{\eta})$ where $\boldsymbol{\eta} = (\eta_{10}, \eta_{01}, \eta_{11})$ and $\eta_{00} = 1 - \mathbf{1}'\boldsymbol{\eta}$ with

$$\eta_a = \pi_{a|00}\theta_{00} + \pi_{a|10}\theta_{10} + \pi_{a|01}\theta_{01} + \pi_{a|11}\theta_{11}, \quad a \in \{00, 10, 01, 11\}.$$

With $\boldsymbol{\pi}_{00} = (\pi_{10|00}, \pi_{01|00}, \pi_{11|00})'$, and

$$\boldsymbol{\Phi} = \begin{pmatrix} \pi_{10|10} - \pi_{10|00} & \pi_{10|01} - \pi_{10|00} & \pi_{10|11} - \pi_{10|00} \\ \pi_{01|10} - \pi_{01|00} & \pi_{01|01} - \pi_{01|00} & \pi_{01|11} - \pi_{01|00} \\ \pi_{11|10} - \pi_{11|00} & \pi_{11|01} - \pi_{11|00} & \pi_{11|11} - \pi_{11|00} \end{pmatrix},$$

the parameter vector for this model can be expressed succinctly as

$$\boldsymbol{\eta} = \boldsymbol{\pi}_{00} + \boldsymbol{\Phi}\boldsymbol{\theta}. \tag{6.1}$$

6.1. Independent Misclassification Errors An alternative, simplified, model to the above assumes there are only four misclassification parameters, specificity and sensitivity for each marginal disease, and that the joint errors can be found assuming independence. Examples of this model can be found in Pfeiffer et al. (2002) and Tebbs et al. (2013), among others. Formally, if $\pi_0^{(i)}$ and $\pi_1^{(i)}$ are the specificity and sensitivity, respectively, for the test under the i th disease, $i = 1, 2$, then we assume $\pi_{10|00} = (1 - \pi_0^{(1)})\pi_0^{(2)}$, $\pi_{10|10} = \pi_1^{(1)}\pi_0^{(2)}$, and so on for all twelve parameters above.

6.2. Identifiability of Model Before addressing the question of unbiased estimation, we first consider conditions to ensure the models presented above are identifiable. This is an important question which has yet to be dealt with explicitly in the literature.

Theorem 4. *Let $\mathbf{Z} \sim IMN_3(c, \boldsymbol{\eta})$ with $\boldsymbol{\eta}$ as in Eq. 6.1. Then, the model is identifiable if and only if the determinant of Φ is non-zero, that is*

$$|\Phi| = \begin{vmatrix} \pi_{10|10} - \pi_{10|00} & \pi_{10|01} - \pi_{10|00} & \pi_{10|11} - \pi_{10|00} \\ \pi_{01|10} - \pi_{01|00} & \pi_{01|01} - \pi_{01|00} & \pi_{01|11} - \pi_{01|00} \\ \pi_{11|10} - \pi_{11|00} & \pi_{11|01} - \pi_{11|00} & \pi_{11|11} - \pi_{11|00} \end{vmatrix} \neq 0.$$

Corollary 3. *For the independent errors model given in Section 6.1, the model is identifiable if and only if both $\pi_0^{(1)} + \pi_1^{(1)} \neq 1$ and $\pi_0^{(2)} + \pi_1^{(2)} \neq 1$.*

Similar to the one-disease case, the conditions of Corollary 3 will always be satisfied if we make the reasonable assumption that all misclassification parameters are greater than 0.5. The more general case, as presented in Theorem 4, is easy to check in a given situation, but does not easily yield itself to simplified conditions.

6.3. Non-Properness of Unbiased Estimator for Two Diseases with Misclassification As in the case with no misclassification, Theorem 1 can be used to construct an unbiased estimator under either of the misclassification models presented above. Since this construction is merely a technical generalization of the previous two cases, it is excluded here.

Likewise, as in the previous cases, we can generalize Theorem 3 to show that this holds under any multinomial sampling plan when misclassification is present. The proof of this result follows directly from Theorem 3 since, assuming the conditions of Theorem 4 hold, $\boldsymbol{\eta}$ is a full rank affine transformation of $\boldsymbol{\theta}$.

Theorem 5. *Let \mathcal{S}_3 be a multinomial sampling plan in four dimensions with boundary points $\mathcal{B}_{\mathcal{S}_3}$ such that at each step the i th coordinate, $i = 0, 1, 2, 3$, is increased with probability η_i , where $\boldsymbol{\eta}$ is as given in Eq. 6.1. Then, there exists no proper unbiased estimator of \mathbf{p} under \mathcal{S}_3 .*

7 Discussion

We have shown that, outside of the standard case when testing one disease with no misclassification, it is impossible to get a proper unbiased estimator in the group testing problem. This result holds very generally, under any design from the classes of binomial or multinomial sampling plans, not only those previously considered in the literature. While, for the multinomial case, we have provided proofs only for two diseases, the same techniques can be applied to show this result holds for any number of traits.

Of course, such scenarios are the norm in applications, so the question remains as to how estimation should best be carried out in light of the

present bias. For one disease with misclassification, there do exist limited results for this problem. For example, under fixed binomial sampling, the first order bias correction given in Tu et al. (1995) can be used to construct a debiased estimator. Still, much more work is needed in this area, analogous to the wide array of estimators in the literature for one disease when no misclassification is assumed. One possible approach is to construct minimal bias estimators as described in Sirazhdinov (1956) and Hall (1963), although approaches minimizing the risk are generally more favored in the statistical literature. Alternatively, approaches such as those found in Bilder and Tebbs (2005) and Hepworth and Biggerstaff (2017), among others, could possibly be extended to include misclassification. More work is needed to understand how such approaches might generalize, and what the properties of the resultant estimators will be.

For the two disease case, unfortunately, it is much more difficult to give recommendations at this time. The problem in this case is much harder since it is both multivariate and has a restricted parameter space, even without misclassification. There currently exist no results in the literature related to bias reduction for the the two disease group testing scenario. This will be an important area for future research if group testing methods are to be applied in such cases.

8 Proofs

8.1. *Proof of Theorem 1* Since h is analytic, both h and $g(\boldsymbol{\mu}) = \frac{h(\boldsymbol{\mu})}{\mu_0^c}$ can be expanded as a Taylor series over an appropriate region, say R . This expansion has the form,

$$g(\boldsymbol{\mu}) = \sum_{x_1, \dots, x_t=0}^{\infty} \frac{1}{x_1! \cdots x_t!} \left. \frac{\partial^{\sum_{i=1}^t x_i} g(\boldsymbol{\mu})}{\partial \mu_1^{x_1} \cdots \partial \mu_t^{x_t}} \right|_{\boldsymbol{\mu}=\mathbf{0}} \prod_{i=1}^t \mu_i^{x_i}.$$

Then, we have

$$\begin{aligned} E(f(\mathbf{X})) &= \sum_{x_1, \dots, x_t=0}^{\infty} f(\mathbf{x}) \binom{c + \sum_{i=1}^t x_i - 1}{c - 1, x_1, \dots, x_t} \mu_0^c \prod_{i=1}^t \mu_i^{x_i} \\ &= \mu_0^c \sum_{x_1, \dots, x_t=0}^{\infty} \frac{1}{x_1! \cdots x_t!} \left. \frac{\partial^{\sum_{i=1}^t x_i} g(\boldsymbol{\mu})}{\partial \mu_1^{x_1} \cdots \partial \mu_t^{x_t}} \right|_{\boldsymbol{\mu}=\mathbf{0}} \prod_{i=1}^t \mu_i^{x_i} \\ &= \mu_0^c g(\boldsymbol{\mu}) \\ &= h(\boldsymbol{\mu}), \end{aligned}$$

for all $\boldsymbol{\mu} \in \text{int}(\Psi)$.

Conversely, if h is estimable unbiasedly, we have for a function $\delta(\mathbf{x})$ and any $\boldsymbol{\mu} \in \text{int}(M)$

$$h(\boldsymbol{\mu}) = \sum_{x_1, \dots, x_t=0}^{\infty} \delta(\mathbf{x}) \binom{c + \sum_{i=1}^t x_i - 1}{c - 1, x_1, \dots, x_t} \mu_0^c \prod_{i=1}^t \mu_i^{x_i}.$$

Since this holds for any $\boldsymbol{\mu} \in \text{int}(\Psi)$, h is an analytic function on R , hence has a unique Taylor expansion. It follows then that

$$\begin{aligned} g(\boldsymbol{\mu}) &= \sum_{x_1, \dots, x_t=0}^{\infty} \frac{1}{x_1! \cdots x_t!} \left. \frac{\partial^{\sum_{i=1}^t x_i} g(\boldsymbol{\mu})}{\partial \mu_1^{x_1} \cdots \partial \mu_t^{x_t}} \right|_{\boldsymbol{\mu}=\mathbf{0}} \prod_{i=1}^t \mu_i^{x_i} \\ &= \sum_{x_1, \dots, x_t=0}^{\infty} \delta(\mathbf{x}) \binom{c + \sum_{i=1}^t x_i - 1}{c - 1, x_1, \dots, x_t} \prod_{i=1}^t \mu_i^{x_i}, \end{aligned}$$

and equating terms yields

$$\delta(\mathbf{x}) = \frac{(c - 1)!}{(c + \sum_{i=1}^t x_i - 1)!} \left. \frac{\partial^{\sum_{i=1}^t x_i} g(\boldsymbol{\mu})}{\partial \mu_1^{x_1} \cdots \partial \mu_t^{x_t}} \right|_{\boldsymbol{\mu}=\mathbf{0}} = f(\mathbf{x})$$

for each \mathbf{x} .

8.2. *Proof of Result 1* To apply Theorem 1 in this case, we will require the following lemma.

Lemma 2. Let $g(\theta) = \frac{(\pi_1 - \theta)^\xi}{(1 - \theta)^c}$, where $\xi = 1/k$. Then, for any non-negative integer t ,

$$\frac{d^t g(\theta)}{d\theta^t} = \sum_{i=0}^t \binom{t}{i} \frac{(\pi_1 - \theta)^{\xi+i-t} (c + i - 1)!}{(1 - \theta)^{c+i} (c - 1)!} \prod_{j=i}^t \frac{(t - j - \xi)}{(t - i - \xi)}.$$

PROOF. For $t = 0$ and $t = 1$ the result is a straightforward calculation, and we prove the general case using induction. Suppose the statement holds for $t = m$ so that

$$\frac{d^{m+1} g(\theta)}{d\theta^{m+1}} = \sum_{i=0}^m \binom{m}{i} \frac{(c + i - 1)!}{(c - 1)!} \prod_{j=i}^m \frac{(m - j - \xi)}{(m - i - \xi)} \frac{d}{d\theta} \left(\frac{(\pi_1 - \theta)^{\xi+i-m}}{(1 - \theta)^{c+i}} \right),$$

and, for each i ,

$$\frac{d}{d\theta} \left(\frac{(\pi_1 - \theta)^{\xi+i-m}}{(1 - \theta)^{c+i}} \right) = \frac{(m - i - \xi)(\pi_1 - \theta)^{\xi+i-m-1}}{(1 - \theta)^{c+i}} + \frac{(c + i)(\pi_1 - \theta)^{\xi+i-m}}{(1 - \theta)^{c+i+1}}.$$

We now look at the resultant coefficients of the terms $\frac{(\pi_1 - \theta)^{\xi+i-(m+1)}}{(1 - \theta)^{c+i}}$ for each i .

For $i = 0$ and $i = m + 1$ we have, respectively, $\prod_{j=0}^{m+1} \frac{(m + 1 - j - \xi)}{(m + 1 - \xi)}$ and $\frac{(c + m + 1 - 1)!}{(c - 1)!}$.

For $1 \leq i \leq m$ we have

$$\begin{aligned} & \binom{m}{i} \frac{(c + i - 1)!}{(c - 1)!} \prod_{j=i}^m \frac{(m - j - \xi)}{(m - i - \xi)} (m - i - \xi) \\ & + \binom{m}{i - 1} \frac{(c + i - 2)!}{(c - 1)!} \prod_{j=i-1}^m \frac{(m - j - \xi)}{(m - i + 1 - \xi)} (c + i - 1) \\ = & \left[\binom{m}{i} + \binom{m}{i - 1} \right] \frac{(c + i - 1)!}{(c - 1)!} \prod_{j=i}^{m+1} \frac{(m + 1 - j - \xi)}{(m + 1 - i - \xi)} \\ = & \binom{m + 1}{i} \frac{(c + i - 1)!}{(c - 1)!} \prod_{j=i}^{m+1} \frac{(m + 1 - j - \xi)}{(m + 1 - i - \xi)}. \end{aligned}$$

Combining yields

$$\frac{d^{m+1}g(\theta)}{d\theta^{m+1}} = \sum_{i=0}^{m+1} \binom{m+1}{i} \frac{(\pi_1 - \theta)^{\xi+i-(m+1)}}{(1 - \theta)^{c+i}} \frac{(c + i - 1)!}{(c - 1)!} \prod_{j=i}^{m+1} \frac{(m + 1 - j - \xi)}{(m + 1 - i - \xi)}.$$

Now, to apply Theorem 1, we have $g(\theta) = \frac{(\pi_1 - \theta)^\xi}{\nu^\xi(1 - \theta)^c}$, with $\xi = 1/k$, which, by Lemma 2 has t th derivative

$$\frac{d^t g(\theta)}{d\theta^t} = \frac{1}{\nu^\xi} \sum_{i=0}^t \binom{t}{i} \frac{(\pi_1 - \theta)^{\xi+i-t}}{(1 - \theta)^{c+i}} \frac{(c + i - 1)!}{(c - 1)!} \prod_{j=i}^t \frac{(t - j - \xi)}{(t - i - \xi)}.$$

Evaluating at $\theta = 0$ yields

$$\left. \frac{d^t g(\theta)}{d\theta^t} \right|_{\theta=0} = \left(\frac{\pi_1}{\nu} \right)^\xi \sum_{i=0}^t \binom{t}{i} \frac{1}{\pi_1^{t-i}} \frac{(c + i - 1)!}{(c - 1)!} \prod_{j=i}^t \frac{(t - j - \xi)}{(t - i - \xi)}, t = 0, 1, 2, \dots$$

Then, direct application of Theorem 1 yields

$$\begin{aligned} \hat{q}_{UB}(y) &= \frac{(c-1)!}{(c+y-1)!} \left(\frac{\pi_1}{\nu}\right)^\xi \sum_{i=0}^y \binom{y}{i} \frac{1}{\pi_1^{y-i}} \frac{(c+i-1)!}{(c-1)!} \prod_{j=i}^y \frac{(y-j-\xi)}{(y-i-\xi)} \\ &= \left(\frac{\pi_1}{\nu}\right)^\xi \sum_{i=0}^y \binom{y}{i} \frac{1}{\pi_1^{y-i}} \frac{(c+i-1)!}{(c+y-1)!} \prod_{j=i}^y \frac{(y-j-\xi)}{(y-i-\xi)}. \end{aligned}$$

Subtracting the above from one gives the desired unbiased estimator of p .

8.3. *Proof of Corollary 1* While it is possible to derive this result algebraically from Result 1, we provide here a much simpler direct proof using Theorem 1. We have $q = h(\theta) = (1-\theta)^\xi$, where $\xi = 1/k$, so we set $g(\theta) = \frac{(1-\theta)^\xi}{(1-\theta)^c} = (1-\theta)^{\xi-c}$. Differentiating t times with respect to θ yields

$$\begin{aligned} g^{(t)}(\theta) &= (-1)^t (\xi - c)(\xi - c - 1) \times \dots \times (\xi - c - t + 1) (1 - \theta)^{\xi - c - t} \\ &= \frac{1}{(c + t - \xi)} \prod_{i=0}^t (c + i - \xi) (1 - \theta)^{\xi - c - i}, t = 0, 1, 2, \dots \end{aligned}$$

Evaluating this derivative at $\theta = 0$ and applying Theorem 1 yields

$$\begin{aligned} \hat{q}_{UB}(y) &= \frac{(c-1)!}{(c+y-1)!} \frac{1}{(c+y-\xi)} \prod_{i=0}^y (c+i-\xi) \\ &= \frac{1}{\left(1 - \frac{1}{k(c+y)}\right)} \prod_{i=0}^y \left(1 - \frac{1}{k(c+i)}\right), y = 0, 1, 2, \dots \end{aligned}$$

As above, the unbiased estimator of p is then found by subtracting this value from 1.

8.4. *Proof of Theorem 2* For each $(x, y) \in \mathcal{B}$, let $K(x, y)$ be the number of ways to reach the given point, and suppose that $f(x, y)$ is an unbiased estimator of $h(\theta) = q$. Then, we have

$$h(\theta) = \sum_{i=0}^{\infty} \frac{1}{i!} \left. \frac{\partial^i h(\theta)}{\partial \theta^i} \right|_{\theta=0} \theta^i = \sum_{(x,y) \in \mathcal{B}} f(x, y) K(x, y) \theta^x (1-\theta)^y \text{ for all } \theta. \tag{8.1}$$

Since the coefficients for each power of θ on both sides of the equality must be the same, there must exist a point $(0, y^*) \in \mathcal{B}$ such that $f(0, y^*) K(0, y^*) = h(0)$. Now, there is at most one path to any point on the y -axis so, since

$(0, y^*) \in \mathcal{B}$, we have $K(0, y^*) = 1$. This yields, $f(0, y^*) = h(0) = \left(\frac{\pi_1}{\pi_1 + \pi_0 - 1}\right)^{1/k} > 1$ whenever $\pi_0 < 1$.

Suppose now that $\pi_0 = 1$ and $\pi_1 < 1$. From the above argument, the term on the right hand side of Eq. 8.1 associated with the point $(0, y^*)$ reduces to $(1 - \theta)^{y^*}$, so that

$$\sum_{(x,y) \in \mathcal{B} \setminus \{(0,y^*)\}} f(x, y)K(x, y)\theta^x(1 - \theta)^y = q - (1 - \theta)^{y^*} \text{ for all } \theta.$$

Allowing $q \rightarrow 0$, which is equivalent to $\theta \rightarrow \pi_1$, we have

$$\sum_{(x,y) \in \mathcal{B} \setminus \{(0,y^*)\}} f(x, y)K(x, y)\theta^x(1 - \theta)^y \rightarrow -(1 - \pi_1)^{y^*} < 0,$$

which implies $f(x, y) < 0$ for at least one point.

8.5. *Proof of Corollary 2* From Eq. 8.1 in the proof of Theorem 2, we see that any sampling plan yielding an unbiased estimator must have exactly one point on the y axis among its boundary points. If $Y \sim IMN_1(c, 1 - \theta)$ is the number of negatives until c positives are observed, however, then sampling stops if and only if a point on the line $x = c$ is reached. This implies that there is no stopping point along the y axis for the random variable Y , hence no unbiased estimator can exist.

8.6. *Proof of Result 2* As in Result 1, to apply Theorem 1 we require the following lemma giving derivatives of the function $g(\theta)$.

Lemma 3. Let $g(\theta) = \frac{h(\theta)}{\theta_{00}^c}$. Then, for non-negative integers z_{10}, z_{01}, z_{11} and $\theta \in \Psi_Z$,

$$(i) \frac{\partial^{z_{10}+z_{01}+z_{11}} g_{00}(\theta)}{\partial \theta_{10}^{z_{10}} \partial \theta_{01}^{z_{01}} \partial \theta_{11}^{z_{11}}} = \prod_{j=0}^{z_{10}+z_{01}+z_{11}} \frac{(c + j - 1/k)}{(c + z_{10} + z_{01} + z_{11} - 1/k)} (1 - \mathbf{1}'\theta)^{1/k - c - z_{10} - z_{01} - z_{11}};$$

$$(ii) \frac{\partial^{z_{10}+z_{01}+z_{11}} g_{10}(\theta)}{\partial \theta_{10}^{z_{10}} \partial \theta_{01}^{z_{01}} \partial \theta_{11}^{z_{11}}} = \frac{(c + z_{10} - 1)!}{(c - 1)!} (1 - \theta_{01} - \theta_{11})^{1/k - z_{01} - z_{11}} \times \sum_{j=0}^{z_{01}+z_{11}} \frac{\theta_{10}^j}{(1 - \mathbf{1}'\theta)^{c+z_{10}+j}} \binom{z_{01} + z_{11}}{j}$$

$$\begin{aligned} &\times \frac{(c + z_{10} + j - 1)!}{(c + z_{10} - 1)!} \\ &\times \prod_{i=j}^{z_{01}+z_{11}} \frac{(c + z_{10} + i - 1/k)}{(c + z_{10} + z_{01} + z_{11} - 1/k)} \\ &\quad - \frac{\partial^{z_{10}+z_{01}+z_{11}} g_{00}(\boldsymbol{\theta})}{\partial \theta_{10}^{z_{10}} \partial \theta_{01}^{z_{01}} \partial \theta_{11}^{z_{11}}}; \end{aligned}$$

$$\begin{aligned} (iii) \quad \frac{\partial^{z_{10}+z_{01}+z_{11}} g_{01}(\boldsymbol{\theta})}{\partial \theta_{10}^{z_{10}} \partial \theta_{01}^{z_{01}} \partial \theta_{11}^{z_{11}}} &= \frac{(c + z_{01} - 1)!}{(c - 1)!} (1 - \theta_{10} - \theta_{11})^{1/k - z_{10} - z_{11}} \\ &\times \sum_{j=0}^{z_{10}+z_{11}} \frac{\theta_{01}^j}{(1 - \mathbf{1}'\boldsymbol{\theta})^{c+z_{01}+j}} \binom{z_{10} + z_{11}}{j} \\ &\times \frac{(c + z_{01} + j - 1)!}{(c + z_{01} - 1)!} \\ &\times \prod_{i=j}^{z_{10}+z_{11}} \frac{(c + z_{01} + i - 1/k)}{(c + z_{10} + z_{01} + z_{11} - 1/k)} \\ &\quad - \frac{\partial^{z_{10}+z_{01}+z_{11}} g_{00}(\boldsymbol{\theta})}{\partial \theta_{10}^{z_{10}} \partial \theta_{01}^{z_{01}} \partial \theta_{11}^{z_{11}}}. \end{aligned}$$

PROOF. Let $\xi = 1/k$. For derivatives of $g_{00}(\boldsymbol{\theta}) = (1 - \mathbf{1}'\boldsymbol{\theta})^{\xi - c}$, we can use the fact that the function is symmetric in θ_{10}, θ_{01} , and θ_{11} and that the partial derivatives can be computed in any order to show the first part iteratively. This is done identically as in the proof of Corollary 1.

For $g_{10}(\boldsymbol{\theta}) = \frac{(1 - \theta_{01} - \theta_{11})^\xi}{\theta_{00}^c} - g_{00}(\boldsymbol{\theta})$, we need only find the partial derivative of the first term. Note that this term is symmetric in θ_{01} and θ_{11} so that the problem is equivalent to finding $\frac{\partial^{z_{10}+r} b(\theta_{10}, \gamma)}{\partial \theta_{10}^{z_{10}} \partial \gamma^r}$, where $b(\theta_{10}, \gamma) = \frac{(1 - \gamma)^\xi}{(1 - \theta_{10} - \gamma)^c}$, which we do by induction.

For the base case, note that for $(z_{10}, \gamma) = (0, 0)$ the result is straightforward. Assume then the result for $(z_{10}, \gamma) = (z, 0)$, so that

$$\frac{\partial^z b(\theta_{10}, \gamma)}{\partial \theta_{10}^z} = \frac{(c + z - 1)!}{(c - 1)!} \frac{(1 - \gamma)^\xi}{(1 - \theta_{10} - \gamma)^{c+z}}.$$

Differentiating with respect to θ_{10} yields

$$\frac{\partial^{z+1} b(\theta_{10}, \gamma)}{\partial \theta_{10}^{z+1}} = \frac{(c + z + 1 - 1)!}{(c - 1)!} \frac{(1 - \gamma)^\xi}{(1 - \theta_{10} - \gamma)^{c+z+1}},$$

so the result holds here as well.

Assume now that the result holds for $(z_{10}, \gamma) = (z_{10}, r)$ so that

$$\begin{aligned} \frac{\partial^{z_{10}+r} b(\theta_{10}, \gamma)}{\partial \theta_{10}^{z_{10}} \partial \gamma^r} &= \frac{(c + z_{10} - 1)!}{(c - 1)!} (1 - \gamma)^{\xi - r} \\ &\times \sum_{j=0}^r \frac{\theta_{10}^j}{(1 - \theta_{10} - \gamma)^{c+z_{10}+j}} \binom{r}{j} \frac{(c + z_{10} + j - 1)!}{(c + z_{10} - 1)!} \\ &\times \prod_{i=j}^r \frac{(c + z_{10} + i - \xi)}{(c + z_{10} + r - \xi)}. \end{aligned}$$

Differentiating with respect to γ yields

$$\begin{aligned} \frac{\partial^{z_{10}+r+1} b(\theta_{10}, \gamma)}{\partial \theta_{10}^{z_{10}} \partial \gamma^{r+1}} &= \frac{(c + z_{10} - 1)!}{(c - 1)!} \sum_{j=0}^r \theta_{10}^j \binom{r}{j} \frac{(c + z_{10} + j - 1)!}{(c + z_{10} - 1)!} \\ &\times \prod_{i=j}^r \frac{(c + z_{10} + i - \xi)}{(c + z_{10} + r - \xi)} \frac{\partial}{\partial \gamma} \left(\frac{(1 - \gamma)^{\xi - r}}{(1 - \theta_{10} - \gamma)^{c+z_{10}+j}} \right), \end{aligned}$$

with

$$\begin{aligned} \frac{\partial}{\partial \gamma} \left(\frac{(1 - \gamma)^{\xi - r}}{(1 - \theta_{10} - \gamma)^{c+z_{10}+j}} \right) &= \frac{(c + z_{10} + j + r - \xi)(1 - \gamma)^{\xi - r - 1}}{(1 - \theta_{10} - \gamma)^{c+z_{10}+j}} \\ &+ \frac{\theta_{10}(c + z_{10} + j)(1 - \gamma)^{\xi - r - 1}}{(1 - \theta_{10} - \gamma)^{c+z_{10}+j+1}}. \end{aligned}$$

For the coefficient of the term $\frac{(1 - \gamma)^{\xi - r - 1}}{(1 - \theta_{10} - \gamma)^{c+z_{10}}}$ this yields

$$\prod_{i=0}^r \frac{(c + z_{10} + i - \xi)}{(c + z_{10} + r - \xi)} (c + z_{10} + r - \xi) = \prod_{i=0}^{r+1} \frac{(c + z_{10} + i - \xi)}{(c + z_{10} + r + 1 - \xi)}.$$

Likewise, for the coefficients of the terms $\frac{(1 - \gamma)^{\xi - r - 1}}{(1 - \theta_{10} - \gamma)^{c+z_{10}+j}}$, $j = 1, 2, \dots, r$, this yields

$$\begin{aligned} &\theta_{10}^j \binom{r}{j} \frac{(c+z_{10}+j-1)!}{(c+z_{10}-1)!} \prod_{i=j}^r \frac{(c+z_{10}+i-\xi)}{(c+z_{10}+r-\xi)} \times (c + z_{10} + j + r - \xi) \\ &+ \theta_{10}^{j-1} \binom{r}{j-1} \frac{(c+z_{10}+j-1-1)!}{(c+z_{10}-1)!} \prod_{i=j-1}^r \frac{(c+z_{10}+i-\xi)}{(c+z_{10}+r-\xi)} \times \theta_{10}(c + z_{10} + j - 1), \end{aligned}$$

which simplifies to

$$\theta_{10}^j \binom{r+1}{j} \frac{(c+z_{10}+j-1)!}{(c+z_{10}-1)!} \prod_{i=j}^r \frac{(c+z_{10}+i-\xi)}{(c+z_{10}+r-\xi)}$$

$$\times \left[\frac{(r+1-j)(c+z_{10}+j+r-\xi)}{r+1} + \frac{j(c+z_{10}+j-1-\xi)}{r+1} \right]$$

and finally to

$$\theta_{10}^j \binom{r+1}{j} \frac{(c+z_{10}+j-1)!}{(c+z_{10}-1)!} \prod_{i=j}^{r+1} \frac{(c+z_{10}+i-\xi)}{(c+z_{10}+r+1-\xi)}.$$

Finally, we have a coefficient for the term $\frac{(1-\gamma)^{\xi-r-1}}{(1-\theta_{10}-\gamma)^{c+z_{10}+r+1}}$

$$\theta_{10}^r \frac{(c+z_{10}+r-1)!}{(c+z_{10}-1)!} \times \theta_{10}(c+z_{10}+r) = \theta_{10}^{r+1} \frac{(c+z_{10}+r+1-1)!}{(c+z_{10}-1)!}.$$

Combining yields the desired result.

Since $b(\theta_{10}, \gamma)$ is a smooth function on the indicated space, the order of derivatives is immaterial and so the above completes the proof by induction. The result for $g_{01}(\boldsymbol{\theta})$ is identical and omitted.

Now, to find an unbiased estimator of $p_{00} = h_{00}(\boldsymbol{\theta})$, we evaluate the derivative from Lemma 3 at $\boldsymbol{\theta} = \mathbf{0}$ and use Theorem 1 to get

$$\begin{aligned} \hat{p}_{00} &= \frac{(c-1)!}{(c+z_{10}+z_{01}+z_{11})!} \prod_{j=0}^{z_{10}+z_{01}+z_{11}} \frac{(c+j-\xi)}{(c+z_{10}+z_{01}+z_{11}-\xi)} \\ &= \frac{(c+z_{10}+z_{01}+z_{11})}{(c+z_{10}+z_{01}+z_{11}-\xi)} \prod_{j=0}^{z_{10}+z_{01}+z_{11}} \frac{(c+j-\xi)}{(c+j)} \\ &= \frac{1}{\left(1 - \frac{\xi}{(c+z_{10}+z_{01}+z_{11})}\right)} \prod_{j=0}^{z_{10}+z_{01}+z_{11}} \left(1 - \frac{\xi}{(c+j)}\right). \end{aligned}$$

The proofs for \hat{p}_{10} and \hat{p}_{01} are nearly identical and are omitted here.

8.7. Proof of Theorem 3 Let $\mathbf{h}^*(\boldsymbol{\theta})$ be the first three components of \mathbf{h} and $\widehat{\mathbf{h}^*(\boldsymbol{\theta})}$ an unbiased estimator under \mathcal{S} . Now, there exist values of $\boldsymbol{\theta} \in \Psi_{\boldsymbol{\theta}}$ such that $\mathbf{1}'\mathbf{h}^*(\boldsymbol{\theta}) > 1$ (for example, $\boldsymbol{\theta} = (.45, .45, .05)$). However, $\mathbf{h}^*(\boldsymbol{\theta})$ is analytic on all values $\Psi_{\boldsymbol{\theta}}$, and so, by the uniqueness of the multivariate Taylor expansion, $E_{\mathcal{S}}(\widehat{\mathbf{h}^*(\boldsymbol{\theta})}) = \mathbf{h}^*(\boldsymbol{\theta})$ for all $\boldsymbol{\theta} \in \Psi_{\boldsymbol{\theta}}$, even those outside of $\Psi_{\mathcal{Z}}$. Then, if $\widehat{\mathbf{h}^*(\boldsymbol{\theta})}$ were a proper estimator under \mathcal{S} , we would have $E_{\mathcal{S}}(\mathbf{1}'\widehat{\mathbf{h}^*(\boldsymbol{\theta})}) \leq 1$ for all $\boldsymbol{\theta}$, so that $\widehat{\mathbf{h}^*(\boldsymbol{\theta})}$ is not unbiased.

8.8. *Proof of Theorem 4* Since the Multinomial is a full rank exponential family (for four class data in three dimensions), it is sufficient to show that there exists a one-to-one mapping from \mathbf{p} to $\boldsymbol{\eta}$. Since, From Lemma 1, $\boldsymbol{\theta}$ is a one-to-one function of \mathbf{p} , we need only show that such a mapping exists from $\boldsymbol{\theta}$ to $\boldsymbol{\eta}$. However, from Eq. 6.1 the correspondence is one-to-one if and only if Φ is non-singular, which is equivalent to the given condition.

8.9. *Proof of Corollary 3* Let $\nu_1 = \pi_0^{(1)} + \pi_1^{(1)} - 1$ and $\nu_2 = \pi_0^{(2)} + \pi_1^{(2)} - 1$. Then, ϕ from Theorem 4 reduces to

$$\begin{vmatrix} \nu_1 \pi_0^{(2)} & -\nu_2(1 - \pi_0^{(1)}) & \nu_1 \pi_0^{(2)} - \nu_2 \pi_1^{(1)} \\ -\nu_1(1 - \pi_0^{(2)}) & \nu_2 \pi_0^{(1)} & \nu_2 \pi_0^{(1)} - \nu_1 \pi_1^{(2)} \\ \nu_1(1 - \pi_0^{(2)}) & \nu_2(1 - \pi_0^{(1)}) & \nu_1(1 - \pi_0^{(2)}) + \nu_2 \pi_1^{(1)} \end{vmatrix} = (\nu_1 \nu_2)^2.$$

Then, $\phi \neq 0$ if and only if both $\nu_1 \neq 0$ and $\nu_2 \neq 0$.

8.10. *Proof of Theorem 5* From Eq. 6.1 we have $\boldsymbol{\theta} = g(\boldsymbol{\eta}) = \Phi^{-1}\boldsymbol{\eta} + \Phi^{-1}\boldsymbol{\pi}_{00}$, so that $\boldsymbol{\theta}$ can be achieved by a full rank affine transformation of $\boldsymbol{\eta}$. But, this implies that the function $\mathbf{p} = h(g(\boldsymbol{\eta}))$ is analytic at the same points $\boldsymbol{\theta} \notin \Psi_Z$ as noted in the proof of Theorem 3. The proof then follows exactly as in the previous theorem.

Acknowledgments. The authors would like to thank the Editor, Associate Editor, and two anonymous referees whose input greatly improved the presentation of this paper.

References

- BILDER, C.R. and TEBBS, J.M. (2005). Empirical Bayes estimation of the disease transmission probability in multiple-vector-transfer designs. *Biom. J.* **47**, 502–516.
- BURROWS, P.M. (1987). Improved estimation of pathogen transmission rates by group testing. *Phytopathology* **77**, 363–365.
- CHIANG, C.L. and REEVES, W.C. (1962). Statistical estimation of virus infection rates in mosquito vector populations. *Am. J. Hyg.* **75**, 377–391.
- DING, J. and XIONG, W. (2015). Robust group testing for multiple traits with misclassification. *J. Appl. Stat.* **42**, 2115–2125.
- DING, J. and XIONG, W. (2016). A new estimator for a population proportion using group testing. *Communications in Statistics – Simulation and Computation* **45**, 101–114.
- DEGROOT, M.H. (1959). Unbiased sequential estimation for binomial populations. *Ann. Math. Stat.* **30**, 80–101.
- DORFMAN, R. (1943). The detection of defective members of large populations. *Ann. Math. Stat.* **14**, 436–440.
- GIBBS, A.J. and GOWER, J.C. (1960). The use of a multiple-transfer method in plant virus transmission studies—some statistical points arising in the analysis of results. *Annals of Applied Biology* **48**, 75–83.
- GIRSHICK, M.A., MOSTELLER, F. and SAVAGE, L. (1946). J Unbiased estimates for certain binomial sampling problems with applications. *Ann. Math. Stat.* **17**, 13–23.

- HABER, G. and MALINOVSKY, Y. (2017). Random walk designs for selecting pool sizes in group testing estimation with small samples. *Biom. J.* **59**, 1382–1398.
- HABER, G., MALINOVSKY, Y. and ALBERT, P.S. (2018). Sequential estimation in the group testing problem. *Seq. Anal.* **37**, 1–17.
- HALL, W.J. (1963). *Estimators with minimum bias*. University of California Press, Berkeley, BELLMAN, R. (ed.), p. 167–199.
- HEPWORTH, G. and BIGGERSTAFF, B. (2017). Bias correction in estimating proportions by pooled testing. *J. Agric. Biol. Environ. Stat.* **22**, 602–614.
- HEPWORTH, G. and WATSON, R. (2009). Debiased estimation of proportions in group testing. *J. R. Stat. Soc. Ser. C* **58**, 105–121.
- HUANG, S., HUANG, M.L., SHEDDEN, K. and WONG, W.K. (2017). Optimal group testing designs for estimating prevalence with uncertain testing errors. *J. R. Stat. Soc. Ser. B* **79**, 1547–1563.
- HUGHES-OLIVER, J.M. and ROSENBERGER, W. (2000). Efficient estimation of the prevalence of multiple rare traits. *Biometrika* **87**, 315–327.
- HUGHES-OLIVER, J.M. and SWALLOW, W.H. (1994). A two-stage adaptive group testing procedure for estimating small proportions. *J. Am. Stat. Assoc.* **89**, 982–993.
- HUNG, M. and SWALLOW, W.H. (1999). Robustness of group testing in the estimation of proportions. *Biometrics* **55**, 231–237.
- KOIKE, K. (1993). Unbiased estimation for sequential multinomial sampling plans. *Seq. Anal.* **12**, 253–259.
- KREMERS, W. (1990). Completeness and unbiased estimation in sequential multinomial sampling. *Seq. Anal.* **9**, 43–58.
- LEHMANN, E.L. and CASELLA, G. (1998). *Theory of point estimation*, 2nd edn. Springer, New York.
- LI, Q., LIU, A. and XIONG, W. (2017). D-optimality of group testing for joint estimation of correlated rare diseases with misclassification. *Stat. Sin.* **27**, 823–838.
- LIU, A., LIU, C., ZHANG, Z. and ALBERT, P.S. (2012). Optimality of group testing in the presence of misclassification. *Biometrika* **99**, 245–251.
- MCAHAN, C.S., TEBBS, J.M. and BILDER, C.R. (2013). Regression models for group testing data with pool dilution effects. *Biostatistics* **14**, 284–298.
- PFEIFFER, R.M., RUTTER, J.L., GAIL, M.H., STRUEWING, J. and GASTWIRTH, J.L. (2002). Efficiency of DNA pooling to estimate joint allele frequencies and measure linkage disequilibrium. *Genet. Epidemiol.* **22**, 94–102.
- SANTOS, J.D. and DORGMAN, D. (2016). An approximate likelihood estimator for the prevalence of infections in vectors using pools of varying sizes. *Biom. J.* **58**, 1248–1256.
- SIRAZHDINOV, S. (1956). On estimators with minimum bias for a binomial distribution. *Theory of Probability and its Applications* **1**, 150–156.
- SWALLOW, W.H. (1985). Group testing for estimating infection rates and probabilities of disease transmission. *Phytopathology* **75**, 882–889.
- TEBBS, J.M., BILDER, C.R. and KOSER, B.K. (2003). An empirical Bayes group-testing approach to estimating small proportions. *Communications in Statistics – Theory and Methods* **32**, 983–995.
- TEBBS, J.M., MCAHAN, C.S. and BILDER, C.R. (2013). Two-stage hierarchical group testing for multiple infections with application to the Infertility Prevention Project. *Biometrics* **69**, 1064–1073.
- THOMPSON, K.H. (1962). Estimation of the proportion of vectors in a natural population of insects. *Biometrics* **18**, 568–578.

- TU, X.M., LITVAK, E. and PAGANO, M. (1995). On the informativeness and accuracy of pooled testing in estimating prevalence of a rare disease: application to HIV screening. *Biometrika* **82**, 287–297.
- WARASI, M.S., TEBBS, J.M., MCMAHAN, C.S. and BILDER, C.R. (2016). Estimating the prevalence of multiple diseases from two-stage hierarchical pooling. *Statistics In Medicine* **35**, 3851–3864.
- ZHANG, Z., LIU, C., KIM, S. and LIU, A. (2014). Prevalence estimation subject to misclassification: the mis-substitution bias and some remedies. *Stat. Med.* **33**, 4482–4500.

GREGORY HABER
YAAKOV MALINOVSKY
DEPARTMENT OF MATHEMATICS AND
STATISTICS, UNIVERSITY OF MARYLAND,
BALTIMORE COUNTY, 1000 HILLTOP
CIRCLE, BALTIMORE, MD, 21250, USA
E-mail: ghaber1@umbc.edu
yaakovm@umbc.edu

Paper received: 7 November 2017.