Web-based Supporting Materials for "Group testing regression models with dilution submodels"

Md S. Warasi,[1] Christopher S. McMahan,[2] Joshua M. Tebbs,[3,*], and Christopher R. Bilder[4]

[1]Department of Mathematics and Statistics, Radford University, Radford, VA 24142, USA
[2]Department of Mathematical Sciences, Clemson University, Clemson, SC 29634, USA
[3]Department of Statistics, University of South Carolina, Columbia, SC 29208, USA
[4]Department of Statistics, University of Nebraska-Lincoln, Lincoln, NE 68583, USA
[*]*email*: tebbs@stat.sc.edu

Web Appendix A: *Dilution submodels.* We presented four dilution submodels

$$h(k, c_j, \lambda) = \text{pr}_\lambda \left(Z_j = 1 \middle| \sum_{i=1}^{c_j} \widetilde{Y}_{ij} = k \right)$$

in Sections 2.2 and 6.2 in the manuscript. In the first three submodels, the notation $\tau(k, c_j) = (k - c_j)/c_j$. The fourth submodel is from Hung and Swallow (1999).

Logistic submodel:

$$h(k, c_j, \lambda) = \frac{\exp\{\lambda\, \tau(k, c_j)\}}{(1/S_e) + \exp\{\lambda\, \tau(k, c_j)\} - 1}$$

Probit submodel:

$$h(k, c_j, \lambda) = \Phi\{\Phi^{-1}(S_e) + \lambda\tau(k, c_j)\}$$

Complementary log-log submodel:

$$h(k, c_j, \lambda) = 1 - \exp(-\exp[\ln\{-\ln(1 - S_e)\} + \lambda\tau(k, c_j)])$$

Hung and Swallow (1999) submodel:

$$h(k, c_j, \lambda) = \frac{S_e k}{k + \lambda(c_j - k)}$$

Each of these submodels satisfies conditions (i) and (ii) stated in Section 2.2 of the manuscript.

Table 1 in the manuscript provides values of the logistic dilution submodel for different values of λ when $S_e = 0.99$ and the master pool size $c_j \in \{5, 10\}$. On the next page, we show this same table for the probit, complementary log-log, and Hung and Swallow (1999) dilution submodels. Each table assumes $S_e = 0.99$ (as in Table 1 in the manuscript). Larger values of λ correspond to more dilution.

Table A.1: Probit dilution submodel.

	λ	$k=1$	2	3	4	5	6	7	8	9	10
	1.0	0.94	0.96	0.97	0.98	0.99					
$c_j = 5$	1.7	0.83	0.90	0.95	0.98	0.99					
	2.4	0.66	0.81	0.91	0.97	0.99					
	1.0	0.92	0.94	0.95	0.96	0.97	0.97	0.98	0.98	0.99	0.99
$c_j = 10$	1.7	0.79	0.83	0.87	0.90	0.93	0.95	0.97	0.98	0.98	0.99
	2.4	0.57	0.66	0.74	0.81	0.87	0.91	0.95	0.97	0.98	0.99

Table A.2: Complementary log-log dilution submodel.

	λ	$k=1$	2	3	4	5	6	7	8	9	10
	0.8	0.91	0.94	0.96	0.98	0.99					
$c_j = 5$	1.2	0.83	0.89	0.94	0.97	0.99					
	1.6	0.72	0.83	0.91	0.96	0.99					
	0.8	0.89	0.91	0.93	0.94	0.95	0.96	0.97	0.98	0.99	0.99
$c_j = 10$	1.2	0.79	0.83	0.86	0.89	0.92	0.94	0.96	0.97	0.98	0.99
	1.6	0.66	0.72	0.78	0.83	0.87	0.91	0.94	0.96	0.98	0.99

Table A.3: Hung and Swallow (1999) dilution submodel.

	λ	$k=1$	2	3	4	5	6	7	8	9	10
	0.03	0.88	0.95	0.97	0.98	0.99					
$c_j = 5$	0.07	0.77	0.90	0.95	0.97	0.99					
	0.11	0.69	0.85	0.92	0.96	0.99					
	0.03	0.78	0.88	0.93	0.95	0.96	0.97	0.98	0.98	0.99	0.99
$c_j = 10$	0.07	0.61	0.77	0.85	0.90	0.93	0.95	0.96	0.97	0.98	0.99
	0.11	0.50	0.69	0.79	0.85	0.89	0.92	0.95	0.96	0.98	0.99

Web Appendix B: *E-step expressions in Section 3.2.* We provide closed-form expressions for $E(I_{jk}|\mathcal{D}_D, \boldsymbol{\theta})$ and $E(\widetilde{Y}_{ij}|\mathcal{D}_D, \boldsymbol{\theta})$, which are used in the EM algorithm described in Section 3.2. We also provide a simulation-based technique to approximate these expectations, which is less time-consuming when the master pool sizes are large; e.g., $c_j \geq 20$.

Closed-form expressions: The conditional expectation of $I_{jk} = \mathcal{I}(\sum_{i=1}^{c_j} \widetilde{Y}_{ij} = k)$ is given by $E(I_{jk}|\mathcal{D}_D, \boldsymbol{\theta}) = \mu_{jk}/\sum_{s=0}^{c_j} \mu_{js}$, where

$$\mu_{js} = \sum_{\widetilde{Y}_{ij}: I_{js}=1} q(s, c_j, \lambda) \prod_{i=1}^{c_j} p_{ij}^{\widetilde{Y}_{ij}} (1-p_{ij})^{1-\widetilde{Y}_{ij}} \left\{ \left(S_e^{\widetilde{Y}_{ij}} \overline{S}_p^{1-\widetilde{Y}_{ij}} \right)^{Y_{ij}} \left(\overline{S}_e^{\widetilde{Y}_{ij}} S_p^{1-\widetilde{Y}_{ij}} \right)^{1-Y_{ij}} \right\}^{Z_j},$$

for $s = 0, 1, 2, ..., c_j$. In the expression above and henceforth, $\overline{S}_e = 1 - S_e$, $\overline{S}_p = 1 - S_p$, and

$$q(k, c_j, \lambda) = \begin{cases} \overline{S}_p^{Z_j} S_p^{1-Z_j}, & k = 0 \\ h(k, c_j, \lambda)^{Z_j} \{1 - h(k, c_j, \lambda)\}^{1-Z_j}, & k > 0. \end{cases}$$

The conditional expectation of $\widetilde{Y}_{ij}$ is $E(\widetilde{Y}_{ij}|\mathcal{D}_D, \boldsymbol{\theta}) = p_{ij}(S_e^{Y_{ij}} \overline{S}_e^{1-Y_{ij}})^{Z_j} \mu_{ij}^* / \sum_{s=0}^{c_j} \mu_{js}$, where μ_{js} is given above and

$$\mu_{ij}^* = \sum_{\widetilde{\mathbf{Y}}_{(i)j}} q(W_{(-i)j} + 1, c_j, \lambda) \prod_{i' \neq i} p_{ij}^{\widetilde{Y}_{ij}} (1-p_{ij})^{1-\widetilde{Y}_{ij}} \left\{ \left(S_e^{\widetilde{Y}_{ij}} \overline{S}_p^{1-\widetilde{Y}_{ij}} \right)^{Y_{ij}} \left(\overline{S}_e^{\widetilde{Y}_{ij}} S_p^{1-\widetilde{Y}_{ij}} \right)^{1-Y_{ij}} \right\}^{Z_j},$$

where $\widetilde{\mathbf{Y}}_{(i)j} = (\widetilde{Y}_{1j}, ..., \widetilde{Y}_{i-1,j}, \widetilde{Y}_{i+1,j}, ..., \widetilde{Y}_{c_j j})'$ is the collection of the true statuses of those individuals in the jth pool except the ith one and $W_{(-i)j} = \sum_{i' \neq i} \widetilde{Y}_{i'j}$.

Approximations: Calculating $E(I_{jk}|\mathcal{D}_D, \boldsymbol{\theta})$ and $E(\widetilde{Y}_{ij}|\mathcal{D}_D, \boldsymbol{\theta})$ can be too computationally intense when the master pool sizes are large. To avoid this problem, one can use a Gibbs sampler to approximate the expectations. In particular, one can sample the individuals' latent statuses from the full conditional distribution $\widetilde{Y}_{ij}|\widetilde{\mathbf{Y}}_{(i)j}, \mathcal{D}_D, \boldsymbol{\theta} \sim \text{Bernoulli}(\zeta_{ij})$, where $\widetilde{\mathbf{Y}}_{(i)j}$ is defined above, $\zeta_{ij} = \zeta_{ij}^{(1)}/(\zeta_{ij}^{(1)} + \zeta_{ij}^{(0)})$, and

$$\begin{aligned} \zeta_{ij}^{(1)} &= p_{ij} q(W_{(-i)j} + 1, c_j, \lambda)(S_e^{Y_{ij}} \overline{S}_e^{1-Y_{ij}})^{Z_j} \\ \zeta_{ij}^{(0)} &= (1 - p_{ij}) q(W_{(-i)j}, c_j, \lambda)(S_p^{1-Y_{ij}} \overline{S}_p^{Y_{ij}})^{1-Z_j}. \end{aligned}$$

Here is a summary of the Gibbs sampling procedure:

1. Initialize $\widetilde{\mathbf{Y}}^{(0)}$, specify the number of Gibbs iterates G, and set $g = 1$.

2. Sample $\widetilde{Y}_{ij}^{(g)}|\widetilde{\mathbf{Y}}_{(i)j}^{(g-1)}, \mathcal{D}_D, \boldsymbol{\theta} \sim \text{Bernoulli}(\zeta_{ij})$, for $i = 1, 2, ..., c_j$ and $j = 1, 2, ..., J$, where

$$\widetilde{\mathbf{Y}}_{(i)j}^{(g-1)} = (\widetilde{Y}_{1j}^{(g)}, ..., \widetilde{Y}_{i-1,j}^{(g)}, \widetilde{Y}_{i+1,j}^{(g-1)}, ..., \widetilde{Y}_{c_j j}^{(g-1)})'.$$

3. Set $\widetilde{\mathbf{Y}}^{(g)} = (\widetilde{\mathbf{Y}}_1^{(g)}, \widetilde{\mathbf{Y}}_2^{(g)}, ..., \widetilde{\mathbf{Y}}_J^{(g)})'$, where $\widetilde{\mathbf{Y}}_j^{(g)} = (\widetilde{Y}_{1j}^{(g)}, \widetilde{Y}_{2j}^{(g)}, ..., \widetilde{Y}_{c_j j}^{(g)})'$ for $j = 1, 2, ..., J$.

4. Repeat Steps 2 and 3 until $g = G$.

Using these samples, one can approximate $E(I_{jk}|\mathcal{D}_D, \boldsymbol{\theta})$ by $G^{-1}\sum_{g=1}^{G} I(\sum_{i=1}^{c_j} \widetilde{Y}_{ij}^{(g)} = k)$ and $E(\widetilde{Y}_{ij}|\mathcal{D}_D, \boldsymbol{\theta})$ by $G^{-1}\sum_{g=1}^{G} \widetilde{Y}_{ij}^{(g)}$. A sufficient number of burn-in draws should be used before retaining samples to make the approximations.

Web Appendix C: *Covariance matrix estimation for Dorfman testing.* The observed information matrix in Section 3.2 (for Dorfman testing) is given by

$$\mathbb{I}(\boldsymbol{\theta}) = -\frac{\partial^2 Q(\boldsymbol{\theta}, \boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} - \text{cov}\left\{ \frac{\partial l_C(\boldsymbol{\theta}|\mathcal{D}_D, \widetilde{\mathbf{Y}})}{\partial \boldsymbol{\theta}} \Big| \mathcal{D}_D, \boldsymbol{\theta} \right\}. \tag{C.1}$$

We provide details on how to estimate $\mathbb{I}(\boldsymbol{\theta})^{-1}$ for logistic regression; i.e., $\text{logit}\{\text{pr}(\widetilde{Y}_{ij} = 1|\mathbf{x}_{ij}) = \mathbf{x}_{ij}'\boldsymbol{\beta}\}$, using the logistic dilution submodel $h(k, c_j, \lambda)$ in Section 2.2. Derivations for other dilution submodels follow similarly and are available from the first author. The components required to compute the first term on the right-hand side of (C.1) are

$$\begin{aligned}
\frac{\partial^2 Q(\boldsymbol{\theta}, \boldsymbol{\theta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'} &= -\sum_{j=1}^{J}\sum_{i=1}^{c_j} p_{ij}(1 - p_{ij})\mathbf{x}_{ij}\mathbf{x}_{ij}' \\
\frac{\partial^2 Q(\boldsymbol{\theta}, \boldsymbol{\theta})}{\partial \lambda^2} &= -\sum_{j=1}^{J}\sum_{k=1}^{c_j} E(I_{jk}|\mathcal{D}_D, \boldsymbol{\theta}) h(k, c_j, \lambda)\left\{1 - h(k, c_j, \lambda)\right\}\tau(k, c_j)^2
\end{aligned}$$

and $\partial^2 Q(\boldsymbol{\theta}, \boldsymbol{\theta})/\partial \boldsymbol{\beta} \partial \lambda = \mathbf{0}$, where recall $\tau(k, c_j) = (k - c_j)/c_j$ and calculating $E(I_{jk}|\mathcal{D}_D, \boldsymbol{\theta})$ is described in Web Appendix B. Calculating the second term on the right-hand side of (C.1) is more complex. Note that

$$\dot{l}_C(\boldsymbol{\theta}) \equiv \frac{\partial l_C(\boldsymbol{\theta}|\mathcal{D}_D, \widetilde{\mathbf{Y}})}{\partial \boldsymbol{\theta}} = \begin{pmatrix} \sum_{j=1}^{J}\sum_{i=1}^{c_j}(\widetilde{Y}_{ij} - p_{ij})\mathbf{x}_{ij} \\ \sum_{j=1}^{J}\sum_{k=1}^{c_j}\{Z_j - h(k, c_j, \lambda)\} I_{jk}\tau(k, c_j) \end{pmatrix}.$$

Calculating the covariance matrix of this quantity is extremely difficult analytically, so we suggest approximating it (at the point of convergence of the MLE) using the Gibbs sampler described in Web Appendix B. Specifically, after setting $\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}$, draw $\widetilde{Y}_{ij}^{(g)}$, for $i = 1, 2, .., c_j$, $j = 1, 2, ..., J$, and $g = 1, 2, ..., G$. From this sample, calculate

$$\dot{l}_C^{(g)}(\boldsymbol{\theta}) = \begin{pmatrix} \sum_{j=1}^{J}\sum_{i=1}^{c_j}(\widetilde{Y}_{ij}^{(g)} - p_{ij})\mathbf{x}_{ij} \\ \sum_{j=1}^{J}\sum_{k=1}^{c_j}\{Z_j - h(k, c_j, \lambda)\} I_{jk}^{(g)}\tau(k, c_j) \end{pmatrix}$$

for $g = 1, 2, ..., G$, where $I_{jk}^{(g)} = \mathcal{I}(\sum_{i=1}^{c_j} \widetilde{Y}_{ij}^{(g)} = k)$. The second term on the right-hand side of (C.1), at the point of convergence of the MLE, can be approximated by using $\mathbf{C}(\widehat{\boldsymbol{\theta}})$, the

sample covariance matrix of $\dot{l}_C^{(1)}(\boldsymbol{\theta}), \dot{l}_C^{(2)}(\boldsymbol{\theta}), ..., \dot{l}_C^{(G)}(\boldsymbol{\theta})$, evaluated at $\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}$. Finally, one can estimate $\mathbb{I}(\boldsymbol{\theta})$ using $\mathbb{I}(\widehat{\boldsymbol{\theta}}) = -\partial^2 Q(\boldsymbol{\theta}, \widehat{\boldsymbol{\theta}})/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}' - \mathbf{C}(\widehat{\boldsymbol{\theta}})$ and $\mathbb{I}(\boldsymbol{\theta})^{-1}$ using $\mathbb{I}(\widehat{\boldsymbol{\theta}})^{-1}$. Recall that in Section 5 of the manuscript, we observed very strong agreement between the sample standard deviation of the maximum likelihood estimates and the averaged standard error, suggesting that this approach to estimate $\mathbb{I}(\boldsymbol{\theta})^{-1}$ is accurate in large samples.

Web Appendix D: *Additional simulation results from Section 5.* The following results assume the population regression model stated in Section 5 in the manuscript.

- Page 6. **Table D.1:** Simulation results for master pool testing (MPT). This is the companion table to Table 2 in the manuscript. The number of individuals is $N = 5000$.
- Page 7. **Table D.2:** Simulation results for master pool testing (MPT). This is the same table as Table D.1, except the number of individuals is $N = 20000$.
- Page 8. **Table D.3:** Simulation results for Dorfman testing (DT) with variable master pool sizes. True dilution submodel: Logistic. Assumed dilution submodel: Logistic; i.e., the dilution submodel is correctly specified.
- Page 9. **Table D.4:** Simulation results for Dorfman testing (DT) with variable master pool sizes. True dilution submodel: Hung and Swallow (1999). Assumed dilution submodel: Hung and Swallow (1999); i.e., the dilution submodel is correctly specified.
- Page 10. **Table D.5:** Simulation results for Dorfman testing (DT) with variable master pool sizes. True dilution submodel: Logistic. Assumed dilution submodel: Hung and Swallow (1999); i.e., the dilution submodel is misspecified.
- Page 11. **Table D.6:** Simulation results for Dorfman testing (DT) with variable master pool sizes. True dilution submodel: Hung and Swallow (1999). Assumed dilution submodel: Logistic; i.e., the dilution submodel is misspecified.
- Page 12. **Table D.7:** Summary of dilution parameter estimates. This table provides summaries of the dilution parameter estimates $\widehat{\lambda}$ from the simulations in Table 2 and Table D.3.
- Page 13. **Table D.8:** LRT for dilution when the submodel is misspecified. This table shows that our LRT (from Section 4) attains the nominal size under each dilution submodel $h(k, c_j, \lambda)$ and has desirable power properties even when the submodel is misspecified.
- Pages 14-15. **Figures D.1-D.2:** ECDF of $B = 500$ simulated values of the LRT statistic T_{LR} under Dorfman testing (DT) when $\lambda = 0$ (no dilution). The ECDF is compared to the CDF of the $\frac{1}{2}\chi_0^2 + \frac{1}{2}\chi_1^2$ mixture. Figure D.1: $c_j = 5$. Figure D.2: $c_j = 10$.
- Pages 16-17. **Figures D.3-D.4:** ECDF of $B = 500$ simulated values of the LRT statistic T_{LR} under Dorfman testing (DT) when $\lambda > 0$ (dilution). The LRT statistic is calculated under both correct and incorrect dilution submodel assumptions. Figure D.3: $c_j = 5$. Figure D.4: $c_j = 10$.

Table D.1: Simulation study for master pool testing (MPT). Results are based on $B = 500$ simulated group testing data sets from the regression model in Section 5 under the logistic dilution submodel in Section 2.2. "Mean" is the average of the 500 maximum likelihood estimates. "Cov" is the empirical coverage probability of nominal 95% Wald confidence intervals (the margin of error is 0.03). "SD" is the sample standard deviation of the 500 estimates and "SE" is the averaged standard error. The level of dilution is controlled by $\lambda \in \{2.6, 3.8, 5.0\}$; see Table 1. Results for random pooling and homogeneous pooling are shown. The true value is $\boldsymbol{\beta} = (-3, 2, 1)'$. The number of individuals is $N = 5000$.

				Model with constant S_e			Model with dilution submodel		
λ	Pool size			$\widehat{\beta}_0$	$\widehat{\beta}_1$	$\widehat{\beta}_2$	$\widehat{\beta}_0$	$\widehat{\beta}_1$	$\widehat{\beta}_2$
2.6	$c_j = 5$	Random	Mean (Cov)	−3.05 (0.96)	1.92 (0.91)	0.93 (0.95)	−2.98 (0.97)	2.07 (0.95)	1.01 (0.97)
			SD (SE)	0.13 (0.13)	0.15 (0.15)	0.25 (0.26)	0.18 (0.17)	0.24 (0.24)	0.29 (0.30)
		Homogeneous	Mean (Cov)	−3.07 (0.90)	1.94 (0.89)	0.96 (0.94)	−2.81 (0.93)	2.02 (0.92)	1.00 (0.95)
			SD (SE)	0.10 (0.10)	0.15 (0.14)	0.21 (0.21)	0.46 (0.32)	0.19 (0.18)	0.23 (0.23)
	$c_j = 10$	Random	Mean (Cov)	−3.05 (0.95)	1.81 (0.87)	0.85 (0.97)	−3.07 (0.97)	2.21 (0.97)	1.03 (0.97)
			SD (SE)	0.18 (0.18)	0.23 (0.23)	0.39 (0.40)	0.26 (0.26)	0.61 (0.49)	0.59 (0.55)
		Homogeneous	Mean (Cov)	−3.13 (0.84)	1.85 (0.80)	0.93 (0.94)	−2.82 (0.93)	2.07 (0.90)	1.05 (0.97)
			SD (SE)	0.09 (0.10)	0.18 (0.18)	0.28 (0.28)	0.44 (0.30)	0.27 (0.26)	0.33 (0.34)
3.8	$c_j = 5$	Random	Mean (Cov)	−3.15 (0.81)	1.82 (0.77)	0.90 (0.94)	−3.01 (0.92)	2.08 (0.94)	1.05 (0.95)
			SD (SE)	0.14 (0.13)	0.15 (0.15)	0.27 (0.26)	0.24 (0.21)	0.29 (0.27)	0.35 (0.32)
		Homogeneous	Mean (Cov)	−3.20 (0.52)	1.88 (0.82)	0.94 (0.94)	−2.71 (0.87)	2.00 (0.92)	1.00 (0.96)
			SD (SE)	0.10 (0.10)	0.14 (0.13)	0.22 (0.21)	0.55 (0.52)	0.20 (0.20)	0.24 (0.24)
	$c_j = 10$	Random	Mean (Cov)	−3.18 (0.85)	1.59 (0.53)	0.79 (0.96)	−3.10 (0.91)	2.29 (0.91)	1.15 (0.96)
			SD (SE)	0.18 (0.17)	0.23 (0.22)	0.45 (0.41)	0.42 (0.36)	0.87 (0.63)	0.86 (0.69)
		Homogeneous	Mean (Cov)	−3.34 (0.08)	1.71 (0.56)	0.86 (0.87)	−2.58 (0.83)	2.03 (0.95)	1.03 (0.94)
			SD (SE)	0.11 (0.11)	0.18 (0.17)	0.31 (0.28)	0.65 (0.53)	0.26 (0.29)	0.37 (0.36)
5.0	$c_j = 5$	Random	Mean (Cov)	−3.35 (0.24)	1.68 (0.46)	0.83 (0.94)	−3.03 (0.89)	2.04 (0.91)	1.02 (0.96)
			SD (SE)	0.13 (0.13)	0.16 (0.16)	0.28 (0.28)	0.31 (0.31)	0.33 (0.30)	0.37 (0.37)
		Homogeneous	Mean (Cov)	−3.44 (0.00)	1.82 (0.69)	0.90 (0.92)	−2.60 (0.86)	1.91 (0.93)	0.95 (0.93)
			SD (SE)	0.11 (0.11)	0.14 (0.14)	0.23 (0.22)	0.66 (0.77)	0.21 (0.24)	0.26 (0.26)
	$c_j = 10$	Random	Mean (Cov)	−3.55 (0.09)	1.45 (0.30)	0.67 (0.98)	−3.09 (0.82)	2.19 (0.90)	1.10 (0.96)
			SD (SE)	0.18 (0.18)	0.22 (0.23)	0.48 (0.47)	0.74 (0.54)	0.84 (0.64)	0.92 (0.77)
		Homogeneous	Mean (Cov)	−3.74 (0.00)	1.64 (0.40)	0.81 (0.90)	−2.65 (0.87)	1.90 (0.95)	0.95 (0.94)
			SD (SE)	0.12 (0.13)	0.17 (0.17)	0.29 (0.28)	0.74 (0.66)	0.26 (0.30)	0.36 (0.36)

Table D.2: Simulation study for master pool testing (MPT). Results are based on $B = 500$ simulated group testing data sets from the regression model in Section 5 under the logistic dilution submodel in Section 2.2. "Mean" is the average of the 500 maximum likelihood estimates. "Cov" is the empirical coverage probability of nominal 95% Wald confidence intervals (the margin of error is 0.03). "SD" is the sample standard deviation of the 500 estimates and "SE" is the averaged standard error. The level of dilution is controlled by $\lambda \in \{2.6, 3.8, 5.0\}$; see Table 1. Results for random pooling and homogeneous pooling are shown. The true value is $\boldsymbol{\beta} = (-3, 2, 1)'$. The number of individuals is $N = 20000$.

				Model with constant S_e			Model with dilution submodel		
λ	Pool size			$\widehat{\beta}_0$	$\widehat{\beta}_1$	$\widehat{\beta}_2$	$\widehat{\beta}_0$	$\widehat{\beta}_1$	$\widehat{\beta}_2$
2.6	$c_j = 5$	Random	Mean (Cov)	−3.05 (0.91)	1.91 (0.79)	0.95 (0.93)	−3.00 (0.96)	2.02 (0.93)	1.01 (0.95)
			SD (SE)	0.06 (0.06)	0.07 (0.08)	0.13 (0.13)	0.08 (0.09)	0.13 (0.13)	0.14 (0.14)
		Homogeneous	Mean (Cov)	−3.07 (0.67)	1.94 (0.83)	0.96 (0.93)	−2.93 (0.92)	2.02 (0.92)	1.00 (0.95)
			SD (SE)	0.05 (0.05)	0.07 (0.07)	0.11 (0.11)	0.24 (0.17)	0.10 (0.11)	0.12 (0.12)
	$c_j = 10$	Random	Mean (Cov)	−3.04 (0.93)	1.81 (0.60)	0.89 (0.92)	−3.03 (0.95)	2.05 (0.94)	1.03 (0.95)
			SD (SE)	0.09 (0.09)	0.12 (0.11)	0.20 (0.19)	0.11 (0.11)	0.22 (0.23)	0.26 (0.24)
		Homogeneous	Mean (Cov)	−3.12 (0.32)	1.84 (0.52)	0.92 (0.90)	−2.93 (0.89)	2.03 (0.89)	1.01 (0.94)
			SD (SE)	0.05 (0.05)	0.10 (0.09)	0.13 (0.14)	0.22 (0.16)	0.17 (0.15)	0.18 (0.17)
3.8	$c_j = 5$	Random	Mean (Cov)	−3.14 (0.43)	1.81 (0.31)	0.90 (0.90)	−3.00 (0.91)	2.02 (0.91)	1.01 (0.96)
			SD (SE)	0.07 (0.06)	0.08 (0.08)	0.13 (0.13)	0.12 (0.11)	0.15 (0.14)	0.16 (0.16)
		Homogeneous	Mean (Cov)	−3.19 (0.01)	1.87 (0.48)	0.93 (0.91)	−2.91 (0.82)	1.99 (0.92)	0.99 (0.94)
			SD (SE)	0.05 (0.05)	0.07 (0.07)	0.11 (0.11)	0.34 (0.30)	0.11 (0.11)	0.12 (0.13)
	$c_j = 10$	Random	Mean (Cov)	−3.16 (0.53)	1.58 (0.04)	0.77 (0.79)	−3.03 (0.92)	2.02 (0.89)	1.01 (0.95)
			SD (SE)	0.09 (0.09)	0.11 (0.11)	0.21 (0.20)	0.15 (0.15)	0.32 (0.29)	0.31 (0.29)
		Homogeneous	Mean (Cov)	−3.32 (0.00)	1.70 (0.08)	0.84 (0.77)	−3.00 (0.84)	1.97 (0.86)	0.99 (0.93)
			SD (SE)	0.05 (0.05)	0.09 (0.08)	0.15 (0.14)	0.25 (0.25)	0.17 (0.16)	0.19 (0.18)
5.0	$c_j = 5$	Random	Mean (Cov)	−3.34 (0.00)	1.67 (0.01)	0.82 (0.78)	−3.00 (0.94)	2.01 (0.94)	1.01 (0.94)
			SD (SE)	0.07 (0.07)	0.08 (0.08)	0.14 (0.14)	0.16 (0.17)	0.17 (0.16)	0.19 (0.18)
		Homogeneous	Mean (Cov)	−3.44 (0.00)	1.82 (0.26)	0.91 (0.86)	−3.04 (0.71)	1.93 (0.87)	0.96 (0.95)
			SD (SE)	0.05 (0.05)	0.07 (0.07)	0.11 (0.11)	0.47 (0.34)	0.12 (0.11)	0.12 (0.13)
	$c_j = 10$	Random	Mean (Cov)	−3.54 (0.00)	1.43 (0.00)	0.69 (0.76)	−3.13 (0.86)	1.99 (0.86)	0.99 (0.94)
			SD (SE)	0.09 (0.09)	0.12 (0.11)	0.23 (0.22)	0.27 (0.26)	0.38 (0.33)	0.37 (0.34)
		Homogeneous	Mean (Cov)	−3.73 (0.00)	1.63 (0.01)	0.80 (0.70)	−3.31 (0.64)	1.88 (0.80)	0.93 (0.91)
			SD (SE)	0.06 (0.06)	0.08 (0.08)	0.14 (0.14)	0.35 (0.30)	0.16 (0.14)	0.18 (0.18)

Table D.3: Simulation study for Dorfman testing (DT) with variable master pool sizes $c_j \in \{5, 6, ..., 10\}$. Results are based on $B = 500$ simulated group testing data sets from the regression model in Section 5 under the logistic dilution submodel in Section 2.2. "Mean" is the average of the 500 maximum likelihood estimates. "Cov" is the empirical coverage probability of nominal 95% Wald confidence intervals (the margin of error is 0.03). "SD" is the sample standard deviation of the 500 estimates and "SE" is the averaged standard error. The level of dilution is controlled by $\lambda \in \{2.6, 3.8, 5.0\}$; see Table 1. Results for random pooling and homogeneous pooling are shown. The true value is $\boldsymbol{\beta} = (-3, 2, 1)'$. The number of individuals is $N = 5000$.

			Model with constant S_e			Model with dilution submodel		
λ			$\widehat{\beta}_0$	$\widehat{\beta}_1$	$\widehat{\beta}_2$	$\widehat{\beta}_0$	$\widehat{\beta}_1$	$\widehat{\beta}_2$
2.6	Random	Mean (Cov)	−3.07 (0.87)	1.97 (0.92)	0.98 (0.97)	−3.00 (0.94)	2.01 (0.94)	1.00 (0.97)
		SD (SE)	0.09 (0.09)	0.10 (0.09)	0.15 (0.15)	0.10 (0.10)	0.10 (0.10)	0.15 (0.16)
	Homogeneous	Mean (Cov)	−3.09 (0.80)	2.05 (0.89)	1.02 (0.96)	−3.00 (0.96)	2.01 (0.94)	1.00 (0.97)
		SD (SE)	0.09 (0.09)	0.10 (0.09)	0.16 (0.15)	0.10 (0.10)	0.10 (0.10)	0.16 (0.16)
3.8	Random	Mean (Cov)	−3.18 (0.51)	1.90 (0.79)	0.94 (0.93)	−3.00 (0.94)	1.99 (0.95)	0.99 (0.96)
		SD (SE)	0.10 (0.09)	0.10 (0.10)	0.16 (0.16)	0.11 (0.10)	0.11 (0.11)	0.17 (0.17)
	Homogeneous	Mean (Cov)	−3.24 (0.23)	2.12 (0.77)	1.05 (0.91)	−3.00 (0.96)	2.00 (0.94)	0.99 (0.97)
		SD (SE)	0.10 (0.09)	0.11 (0.10)	0.18 (0.16)	0.11 (0.11)	0.10 (0.10)	0.16 (0.16)
5.0	Random	Mean (Cov)	−3.43 (0.00)	1.82 (0.55)	0.88 (0.91)	−3.00 (0.94)	2.00 (0.96)	0.99 (0.95)
		SD (SE)	0.10 (0.10)	0.10 (0.10)	0.18 (0.17)	0.12 (0.12)	0.12 (0.12)	0.19 (0.19)
	Homogeneous	Mean (Cov)	−3.59 (0.00)	2.30 (0.26)	1.14 (0.81)	−3.02 (0.93)	2.01 (0.94)	1.00 (0.95)
		SD (SE)	0.12 (0.11)	0.13 (0.11)	0.22 (0.18)	0.14 (0.13)	0.12 (0.12)	0.18 (0.18)

Note: The logistic submodel has been correctly specified.

Table D.4: Simulation study for Dorfman testing (DT) with variable master pool sizes $c_j \in \{5, 6, ..., 10\}$. Results are based on $B = 500$ simulated group testing data sets from the regression model in Section 5 under the Hung and Swallow (1999) dilution submodel in Appendix A. "Mean" is the average of the 500 maximum likelihood estimates. "Cov" is the empirical coverage probability of nominal 95% Wald confidence intervals (the margin of error is 0.03). "SD" is the sample standard deviation of the 500 estimates and "SE" is the averaged standard error. The level of dilution is controlled by $\lambda \in \{0.03, 0.07, 0.11\}$; see Table A.3. Results for random pooling and homogeneous pooling are shown. The true value is $\boldsymbol{\beta} = (-3, 2, 1)'$. The number of individuals is $N = 5000$.

			Model with constant S_e			Model with dilution submodel		
λ			$\widehat{\beta}_0$	$\widehat{\beta}_1$	$\widehat{\beta}_2$	$\widehat{\beta}_0$	$\widehat{\beta}_1$	$\widehat{\beta}_2$
0.03	Random	Mean (Cov)	−3.13 (0.71)	1.93 (0.87)	0.96 (0.94)	−3.00 (0.93)	2.00 (0.95)	1.00 (0.95)
		SD (SE)	0.10 (0.09)	0.10 (0.10)	0.16 (0.15)	0.10 (0.10)	0.11 (0.10)	0.16 (0.16)
	Homogeneous	Mean (Cov)	−3.18 (0.49)	2.10 (0.83)	1.05 (0.89)	−3.01 (0.94)	2.00 (0.94)	1.01 (0.93)
		SD (SE)	0.10 (0.09)	0.10 (0.10)	0.18 (0.16)	0.11 (0.11)	0.10 (0.11)	0.17 (0.16)
0.07	Random	Mean (Cov)	−3.28 (0.16)	1.88 (0.77)	0.92 (0.94)	−3.00 (0.97)	2.01 (0.95)	1.00 (0.97)
		SD (SE)	0.10 (0.10)	0.09 (0.10)	0.16 (0.16)	0.10 (0.11)	0.11 (0.11)	0.17 (0.18)
	Homogeneous	Mean (Cov)	−3.38 (0.03)	2.20 (0.52)	1.08 (0.86)	−3.01 (0.95)	2.00 (0.94)	0.99 (0.94)
		SD (SE)	0.11 (0.10)	0.12 (0.10)	0.20 (0.17)	0.11 (0.12)	0.11 (0.11)	0.17 (0.17)
0.11	Random	Mean (Cov)	−3.40 (0.02)	1.84 (0.66)	0.90 (0.91)	−3.01 (0.94)	2.01 (0.94)	1.00 (0.95)
		SD (SE)	0.11 (0.10)	0.10 (0.10)	0.17 (0.17)	0.12 (0.11)	0.12 (0.12)	0.18 (0.19)
	Homogeneous	Mean (Cov)	−3.54 (0.00)	2.29 (0.26)	1.15 (0.78)	−3.01 (0.95)	2.01 (0.96)	1.01 (0.95)
		SD (SE)	0.12 (0.11)	0.12 (0.11)	0.22 (0.17)	0.12 (0.12)	0.11 (0.12)	0.18 (0.17)

Note: The Hung and Swallow (1999) submodel has been correctly specified.

Table D.5: Simulation study for Dorfman testing (DT) with variable master pool sizes $c_j \in \{5, 6, ..., 10\}$. Results are based on $B = 500$ simulated group testing data sets from the regression model in Section 5. The true dilution submodel is logistic; the assumed submodel for estimation is Hung and Swallow (1999); i.e., the dilution submodel has been misspecified. "Mean" is the average of the 500 maximum likelihood estimates. "Cov" is the empirical coverage probability of nominal 95% Wald confidence intervals (the margin of error is 0.03). "SD" is the sample standard deviation of the 500 estimates and "SE" is the averaged standard error. The level of dilution is controlled by $\lambda \in \{2.6, 3.8, 5.0\}$; see Table 1. Results for random pooling and homogeneous pooling are shown. The true value is $\boldsymbol{\beta} = (-3, 2, 1)'$. The number of individuals is $N = 5000$.

			Model with constant S_e			Misspecified dilution submodel		
λ			$\widehat{\beta}_0$	$\widehat{\beta}_1$	$\widehat{\beta}_2$	$\widehat{\beta}_0$	$\widehat{\beta}_1$	$\widehat{\beta}_2$
2.6	Random	Mean (Cov)	−3.07 (0.87)	1.97 (0.92)	0.98 (0.97)	−3.01 (0.95)	2.01 (0.94)	1.00 (0.96)
		SD (SE)	0.09 (0.09)	0.10 (0.09)	0.15 (0.15)	0.10 (0.10)	0.10 (0.10)	0.15 (0.16)
	Homogeneous	Mean (Cov)	−3.09 (0.80)	2.05 (0.89)	1.02 (0.96)	−3.00 (0.95)	2.00 (0.95)	1.00 (0.96)
		SD (SE)	0.09 (0.09)	0.10 (0.09)	0.16 (0.15)	0.10 (0.10)	0.10 (0.10)	0.15 (0.16)
3.8	Random	Mean (Cov)	−3.18 (0.51)	1.90 (0.79)	0.94 (0.93)	−3.02 (0.94)	1.98 (0.96)	0.99 (0.95)
		SD (SE)	0.10 (0.09)	0.10 (0.10)	0.16 (0.16)	0.11 (0.10)	0.11 (0.10)	0.17 (0.17)
	Homogeneous	Mean (Cov)	−3.24 (0.23)	2.12 (0.77)	1.05 (0.91)	−3.02 (0.95)	2.01 (0.96)	1.00 (0.96)
		SD (SE)	0.10 (0.09)	0.11 (0.10)	0.18 (0.16)	0.11 (0.11)	0.10 (0.11)	0.16 (0.16)
5.0	Random	Mean (Cov)	−3.43 (0.00)	1.82 (0.55)	0.88 (0.91)	−3.02 (0.94)	2.00 (0.95)	0.98 (0.95)
		SD (SE)	0.10 (0.10)	0.10 (0.10)	0.18 (0.17)	0.12 (0.12)	0.12 (0.12)	0.19 (0.19)
	Homogeneous	Mean (Cov)	−3.59 (0.00)	2.30 (0.26)	1.14 (0.81)	−3.03 (0.95)	2.02 (0.95)	1.01 (0.95)
		SD (SE)	0.12 (0.11)	0.13 (0.11)	0.22 (0.18)	0.12 (0.13)	0.11 (0.12)	0.18 (0.18)

Table D.6: Simulation study for Dorfman testing (DT) with variable master pool sizes $c_j \in \{5, 6, ..., 10\}$. Results are based on $B = 500$ simulated group testing data sets from the regression model in Section 5. The true dilution submodel is Hung and Swallow (1999); the assumed submodel for estimation is logistic; i.e. the dilution submodel has been misspecified. "Mean" is the average of the 500 maximum likelihood estimates. "Cov" is the empirical coverage probability of nominal 95% Wald confidence intervals (the margin of error is 0.03). "SD" is the sample standard deviation of the 500 estimates and "SE" is the averaged standard error. The level of dilution is controlled by $\lambda \in \{0.03, 0.07, 0.11\}$; see Table A.3. Results for random pooling and homogeneous pooling are shown. The true value is $\boldsymbol{\beta} = (-3, 2, 1)'$. The number of individuals is $N = 5000$.

			Model with constant S_e			Misspecified dilution submodel		
λ			$\widehat{\beta}_0$	$\widehat{\beta}_1$	$\widehat{\beta}_2$	$\widehat{\beta}_0$	$\widehat{\beta}_1$	$\widehat{\beta}_2$
0.03	Random	Mean (Cov)	−3.13 (0.70)	1.94 (0.87)	0.96 (0.94)	−2.99 (0.92)	2.01 (0.95)	1.00 (0.95)
		SD (SE)	0.10 (0.09)	0.10 (0.10)	0.16 (0.15)	0.11 (0.10)	0.11 (0.10)	0.16 (0.16)
	Homogeneous	Mean (Cov)	−3.18 (0.52)	2.10 (0.80)	1.04 (0.90)	−3.00 (0.93)	2.01 (0.93)	1.00 (0.93)
		SD (SE)	0.10 (0.09)	0.11 (0.10)	0.18 (0.16)	0.11 (0.11)	0.11 (0.10)	0.17 (0.16)
0.07	Random	Mean (Cov)	−3.28 (0.16)	1.88 (0.77)	0.92 (0.94)	−2.98 (0.96)	2.02 (0.94)	1.00 (0.96)
		SD (SE)	0.10 (0.10)	0.09 (0.10)	0.16 (0.16)	0.10 (0.11)	0.11 (0.11)	0.17 (0.18)
	Homogeneous	Mean (Cov)	−3.38 (0.03)	2.20 (0.51)	1.08 (0.87)	−3.00 (0.94)	2.00 (0.93)	0.98 (0.94)
		SD (SE)	0.11 (0.10)	0.12 (0.10)	0.20 (0.17)	0.12 (0.12)	0.11 (0.11)	0.17 (0.17)
0.11	Random	Mean (Cov)	−3.40 (0.02)	1.84 (0.66)	0.90 (0.91)	−2.99 (0.93)	2.02 (0.94)	1.01 (0.95)
		SD (SE)	0.11 (0.10)	0.10 (0.10)	0.17 (0.17)	0.12 (0.11)	0.12 (0.12)	0.18 (0.19)
	Homogeneous	Mean (Cov)	−3.54 (0.00)	2.29 (0.27)	1.15 (0.77)	−2.99 (0.95)	1.99 (0.94)	1.00 (0.95)
		SD (SE)	0.12 (0.11)	0.12 (0.11)	0.22 (0.17)	0.13 (0.12)	0.11 (0.12)	0.18 (0.17)

Table D.7: Simulation results for the dilution parameter λ under Dorfman testing (DT). Equal master pool sizes results ($c_j = 5$ and $c_j = 10$) pertain to Table 1; unequal master pool results pertain to Table D.3. "Mean" is the average of the $B = 500$ maximum likelihood estimates. "Cov" is the empirical coverage probability of nominal 95% Wald confidence intervals (the margin of error is 0.03). "SD" is the sample standard deviation of the 500 estimates and "SE" is the averaged standard error. The level of dilution is controlled by $\lambda \in \{2.6, 3.8, 5.0\}$; see Table 1.

		Random pooling			Homogeneous pooling		
c_j		$\lambda = 2.6$	3.8	5.0	$\lambda = 2.6$	3.8	5.0
5	Mean (Cov)	2.33 (0.79)	3.69 (0.95)	5.00 (0.94)	2.31 (0.82)	3.69 (0.97)	4.98 (0.97)
	SD (SE)	1.15 (1.03)	0.64 (0.54)	0.27 (0.27)	1.16 (1.02)	0.63 (0.58)	0.32 (0.32)
10	Mean (Cov)	2.50 (0.92)	3.77 (0.97)	4.99 (0.96)	2.38 (0.90)	3.75 (0.97)	4.98 (0.97)
	SD (SE)	0.68 (0.60)	0.29 (0.28)	0.17 (0.18)	0.88 (0.77)	0.43 (0.39)	0.25 (0.25)
Unequal	Mean (Cov)	2.54 (0.95)	3.77 (0.95)	4.99 (0.97)	2.48 (0.95)	3.76 (0.94)	4.98 (0.96)
	SD (SE)	0.63 (0.79)	0.35 (0.33)	0.19 (0.20)	0.72 (0.86)	0.43 (0.40)	0.26 (0.26)

Note: We occasionally experienced numerical difficulties when estimating the (large-sample) variance of $\widehat{\lambda}$ at the lowest level of dilution (i.e., only when $\lambda = 2.6$). Specifically, estimated (large-sample) standard errors for $\widehat{\lambda}$ were negative in about 3% of the simulations and were therefore removed.

Table D.8: LRT for dilution when the submodel is misspecified. Results are based on $B = 500$ simulated group testing data sets from the regression model in Section 5 while incorrectly assuming the logistic dilution submodel in Section 2.2. The true submodel is either the submodel specified in Hung and Swallow, HS (1999), the probit, or the complementary log-log (CL). The margin of error of the estimated size when $\lambda = 0$, assuming a 99% confidence level, is 0.03. For each submodel, the amount of dilution increases as λ increases. The results for master pool testing (MPT) and Dorfman testing (DT) are shown.

		HS					Probit					CL				
Pool size		$\lambda = 0$	0.02	0.04	0.06	0.08	$\lambda = 0$	0.5	1.0	1.5	2.0	$\lambda = 0$	0.4	0.8	1.2	1.6
							Random pooling									
$c_j = 5$	MPT	0.05	0.11	0.19	0.24	0.31	0.06	0.05	0.09	0.18	0.27	0.05	0.06	0.14	0.20	0.30
	DT	0.05	0.37	0.78	0.94	0.99	0.04	0.09	0.27	0.66	0.97	0.04	0.10	0.38	0.90	1.00
$c_j = 10$	MPT	0.06	0.18	0.22	0.24	0.29	0.06	0.08	0.14	0.22	0.26	0.05	0.11	0.17	0.23	0.26
	DT	0.04	0.94	1.00	1.00	1.00	0.05	0.10	0.51	0.99	1.00	0.04	0.19	0.78	1.00	1.00
							Homogeneous pooling									
$c_j = 5$	MPT	0.05	0.10	0.16	0.17	0.19	0.05	0.07	0.10	0.15	0.26	0.05	0.07	0.15	0.19	0.27
	DT	0.07	0.34	0.78	0.96	0.99	0.04	0.12	0.28	0.74	0.99	0.04	0.15	0.50	0.92	1.00
$c_j = 10$	MPT	0.05	0.23	0.28	0.29	0.36	0.03	0.12	0.25	0.39	0.44	0.06	0.16	0.35	0.49	0.44
	DT	0.05	0.73	0.99	1.00	1.00	0.04	0.18	0.51	0.96	1.00	0.05	0.21	0.73	0.99	1.00

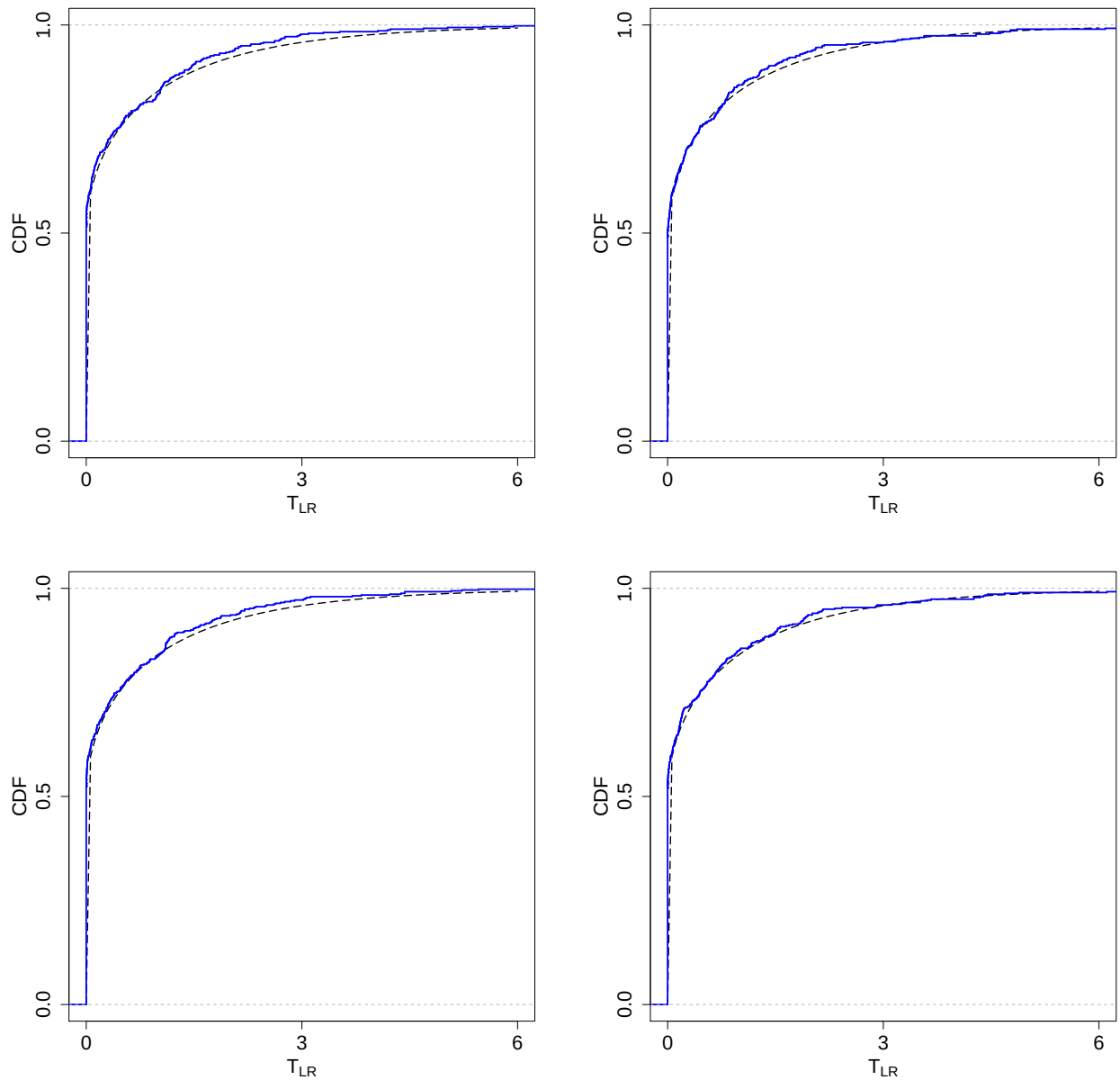

Figure D.1: Empirical CDF (ECDF) of $B = 500$ simulated values of the LRT statistic T_{LR} under Dorfman testing (DT) when $\lambda = 0$; i.e., $H_0 : \lambda = 0$ is true. The ECDF is shown in blue. The CDF of the $\frac{1}{2}\chi_0^2 + \frac{1}{2}\chi_1^2$ mixture is shown dashed in black. Top: Hung and Swallow (1999) dilution submodel. Bottom: Logistic dilution submodel. Left: Random pooling. Right: Homogeneous pooling. For each of the $B = 500$ data sets, the number of individuals is $N = 5000$ and the master pool size is $c_j = 5$.

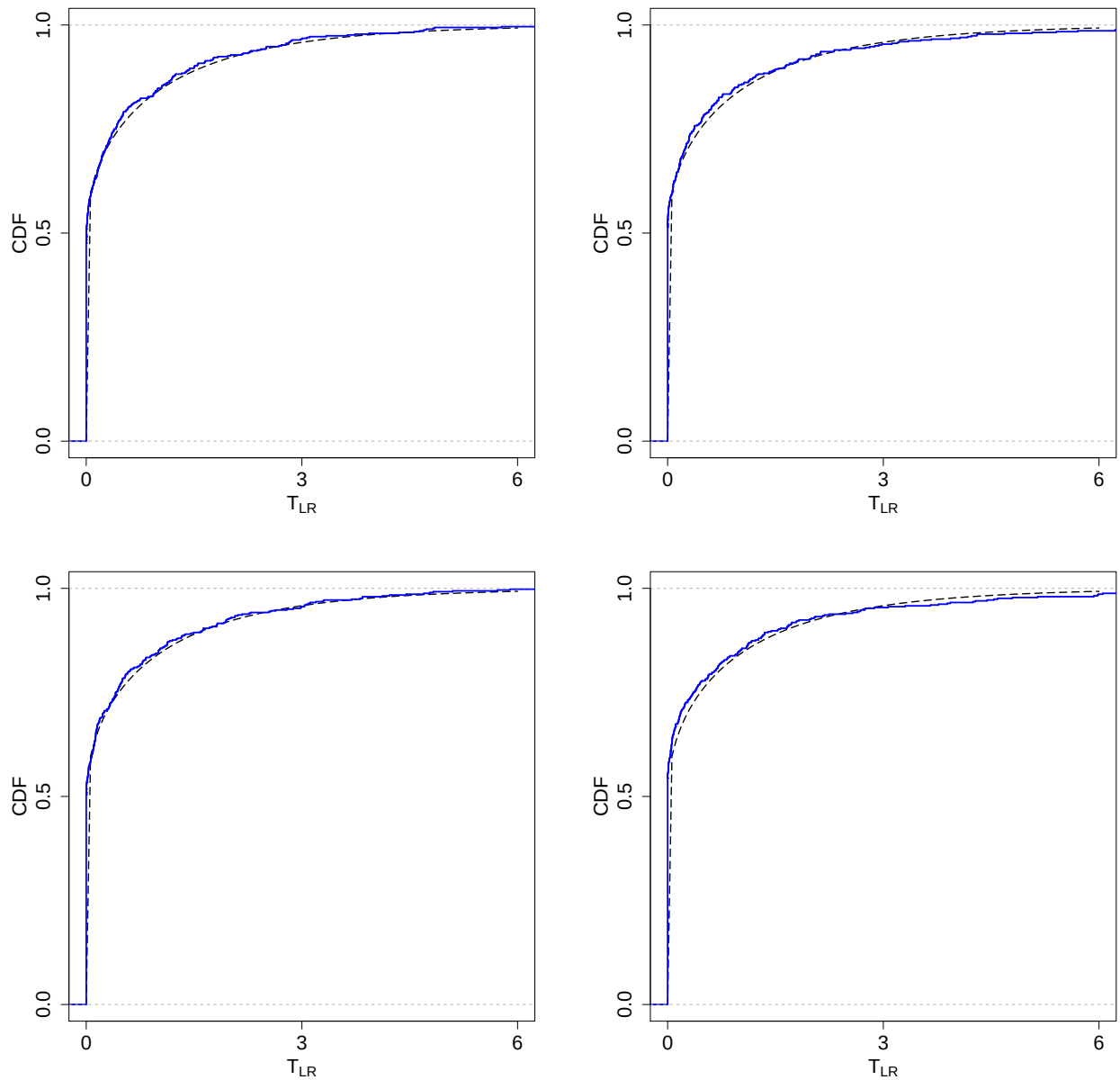

Figure D.2: Empirical CDF (ECDF) of $B = 500$ simulated values of the LRT statistic T_{LR} under Dorfman testing (DT) when $\lambda = 0$; i.e., $H_0 : \lambda = 0$ is true. The ECDF is shown in blue. The CDF of the $\frac{1}{2}\chi_0^2 + \frac{1}{2}\chi_1^2$ mixture is shown dashed in black. Top: Hung and Swallow (1999) dilution submodel. Bottom: Logistic dilution submodel. Left: Random pooling. Right: Homogeneous pooling. For each of the $B = 500$ data sets, the number of individuals is $N = 5000$ and the master pool size is $c_j = 10$.

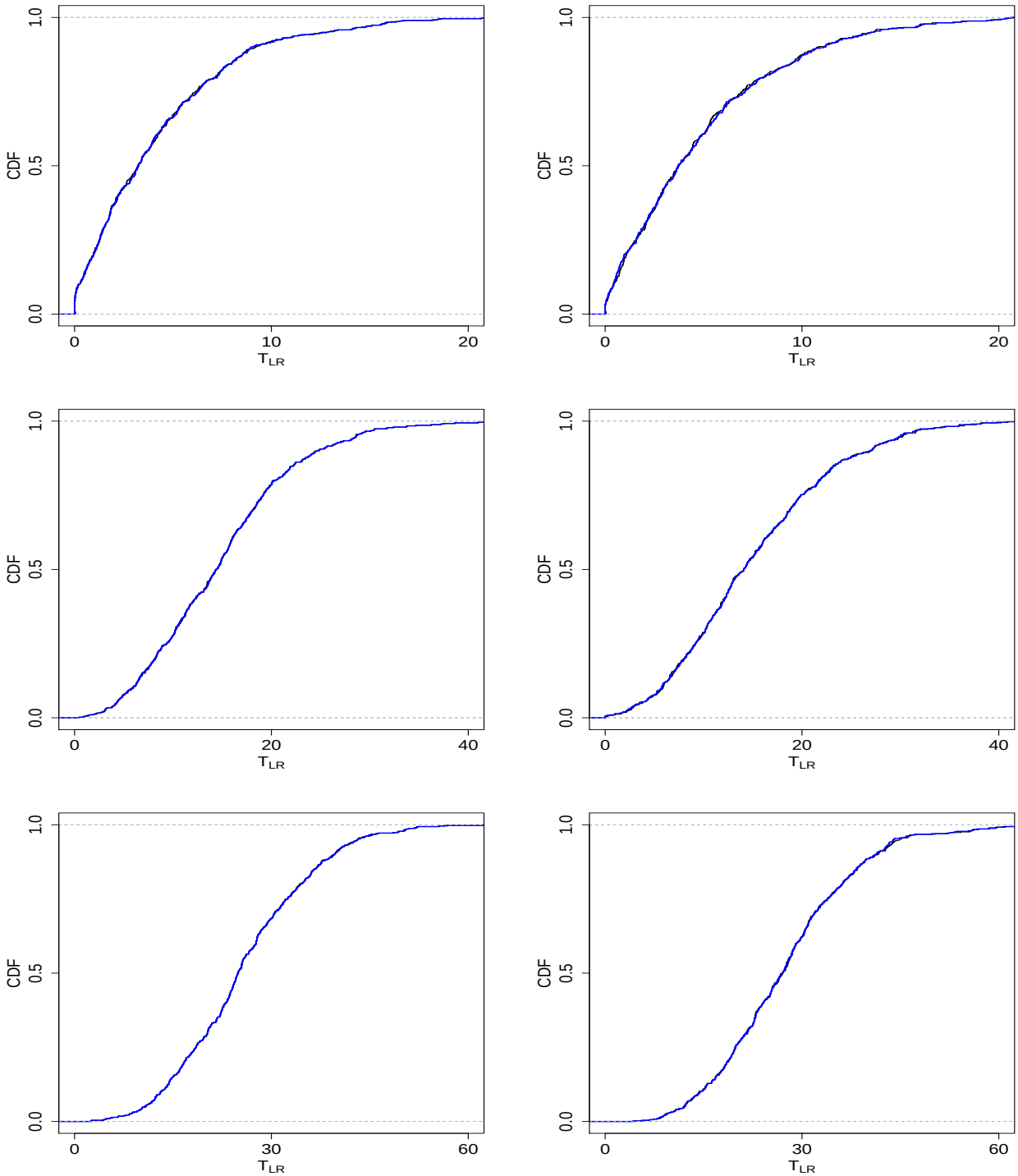

Figure D.3: Empirical CDF (ECDF) of $B = 500$ simulated values of the LRT statistic T_{LR} under Dorfman testing (DT) when $\lambda > 0$; i.e., $H_0 : \lambda = 0$ is not true. The Hung and Swallow (1999) dilution submodel is the true submodel. Results for Hung and Swallow (1999) are shown dashed in black; results under the misspecified logistic submodel are shown in blue. Top: $\lambda = 0.03$. Middle: $\lambda = 0.07$. Bottom: $\lambda = 0.11$. Left: Random pooling. Right: Homogeneous pooling. For each of the $B = 500$ data sets, the number of individuals is $N = 5000$ and the master pool size is $c_j = 5$.

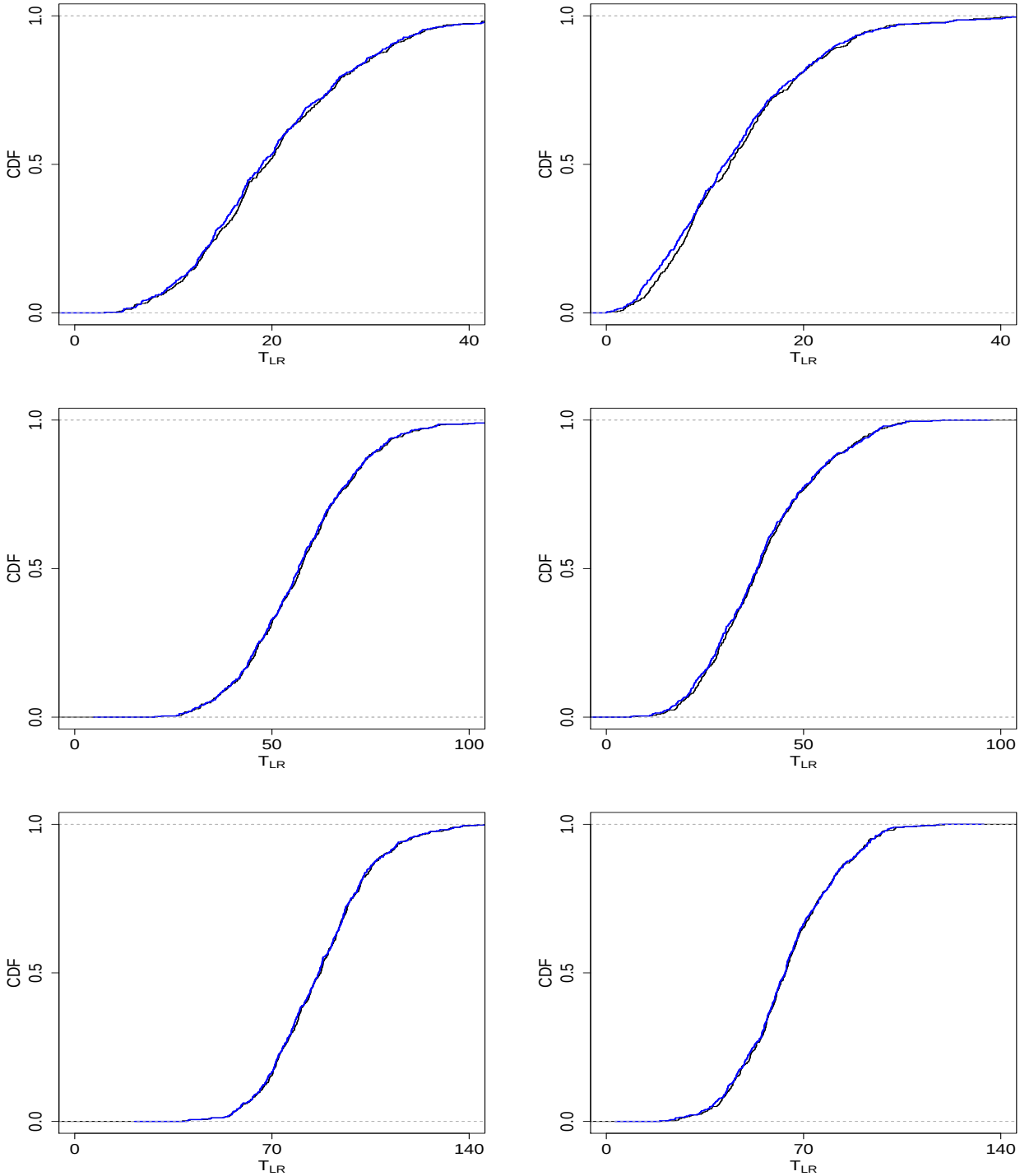

Figure D.4: Empirical CDF (ECDF) of $B = 500$ simulated values of the LRT statistic T_{LR} under Dorfman testing (DT) when $\lambda > 0$; i.e., $H_0 : \lambda = 0$ is not true. The Hung and Swallow (1999) dilution submodel is the true submodel. Results for Hung and Swallow (1999) are shown dashed in black; results under the misspecified logistic submodel are shown in blue. Top: $\lambda = 0.03$. Middle: $\lambda = 0.07$. Bottom: $\lambda = 0.11$. Left: Random pooling. Right: Homogeneous pooling. For each of the $B = 500$ data sets, the number of individuals is $N = 5000$ and the master pool size is $c_j = 10$.