

Reducing the Effects of Careless Responses on Item Calibration in Item Response Theory

Jeffrey M. Patton,
Ying Cheng, & Ke-Hai Yuan
University of Notre Dame
<http://irtND.wikispaces.com>

Qi Diao
CTB/McGraw-Hill



Prevalence of Carelessness


- In psychological and survey research, the prevalence of careless responses from unmotivated participants has been repeatedly reported. Over 50% of examinees responded to one or more items carelessly (Baer, Ballenger, Berry, & Wetter, 1997; Berry et al., 1992).
- In low-stakes educational testing, the same problem persists due to low motivation. For NAEP, 45% of grade 12 students reported that they did not try as hard on the math NAEP test as they did on other math tests taken in school that year, according to the NSF website.

Consequences of Careless Responses in Pretesting

- Consequences:
 - biased item parameter estimates (Nering, 1998; Oshima, 1994; Wise et al., 2004)
 - biased item and test information functions (van Barneveld, 2007)
 - biased ability estimates (De Ayala et al., 2001; Meijer & Sijtsma, 2001)

[Goals]

- What are the effects of careless responses on item parameter estimates?
- Can we reduce these effects by statistically detecting & removing careless responders?



Study 1

What are the effects of careless responses on item parameter estimates?

Simulation Scenario

- Administer a 40-item psychological questionnaire (e.g., a depression scale) to 500 subjects
- Each question has two options: agree or disagree
- Subjects exhibit one of two response styles:
 - “normal” responders: the probability of agreeing is given by the 2PLM:

$$P(u_j|\theta, \gamma_j) = \frac{e^{a_j(\theta-b_j) \cdot u_j}}{1 + e^{a_j(\theta-b_j)}}$$

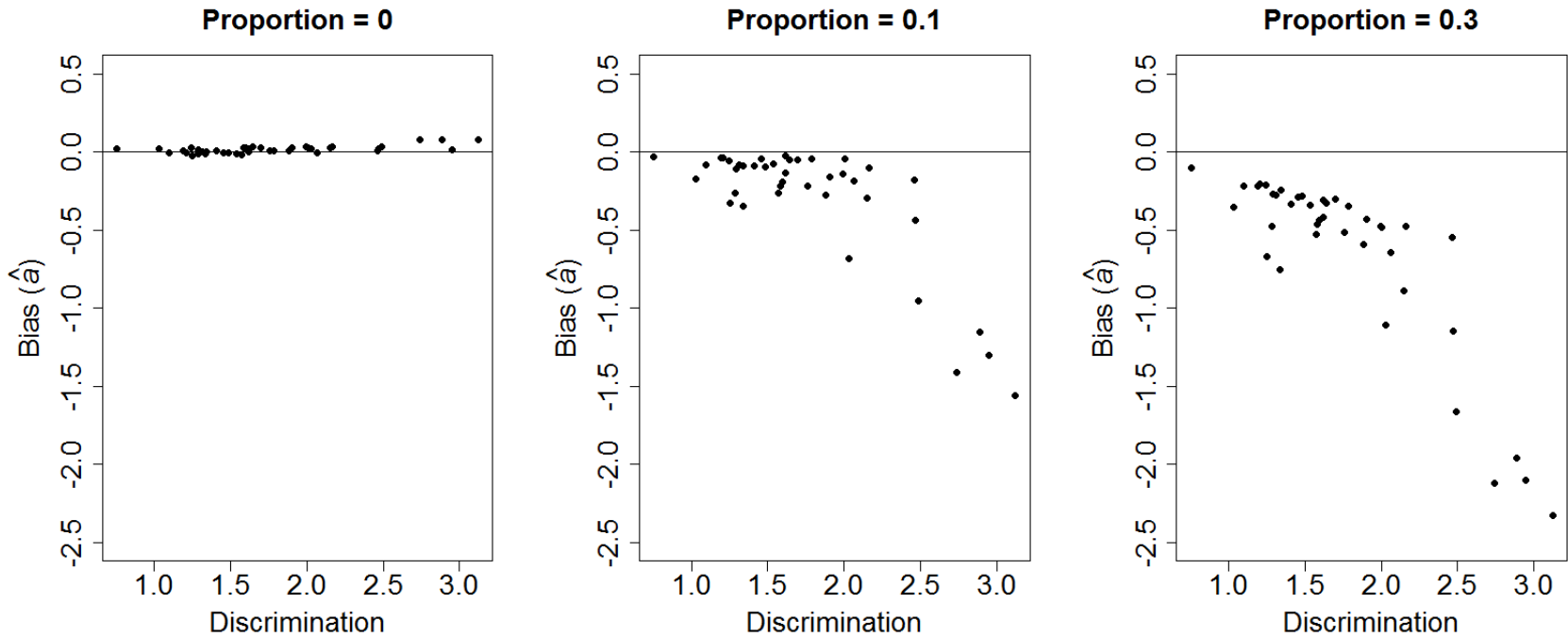
- “careless” responders: randomly choose a response for each item
- A random subsample of $p = 0\%$, 10% , or 30% of the 500 subjects are chosen to be careless

[Method]

- Draw 500 ability parameters from $N(0,1)$
- A percentage p of these subjects are chosen to be careless (thus, true ability and response style are independent)
- Generate data, fit the 2PLM
- Repeat 100 times; yields 100 sets of item parameter estimates for each condition
- Outcomes: bias & SEs of item parameter estimates

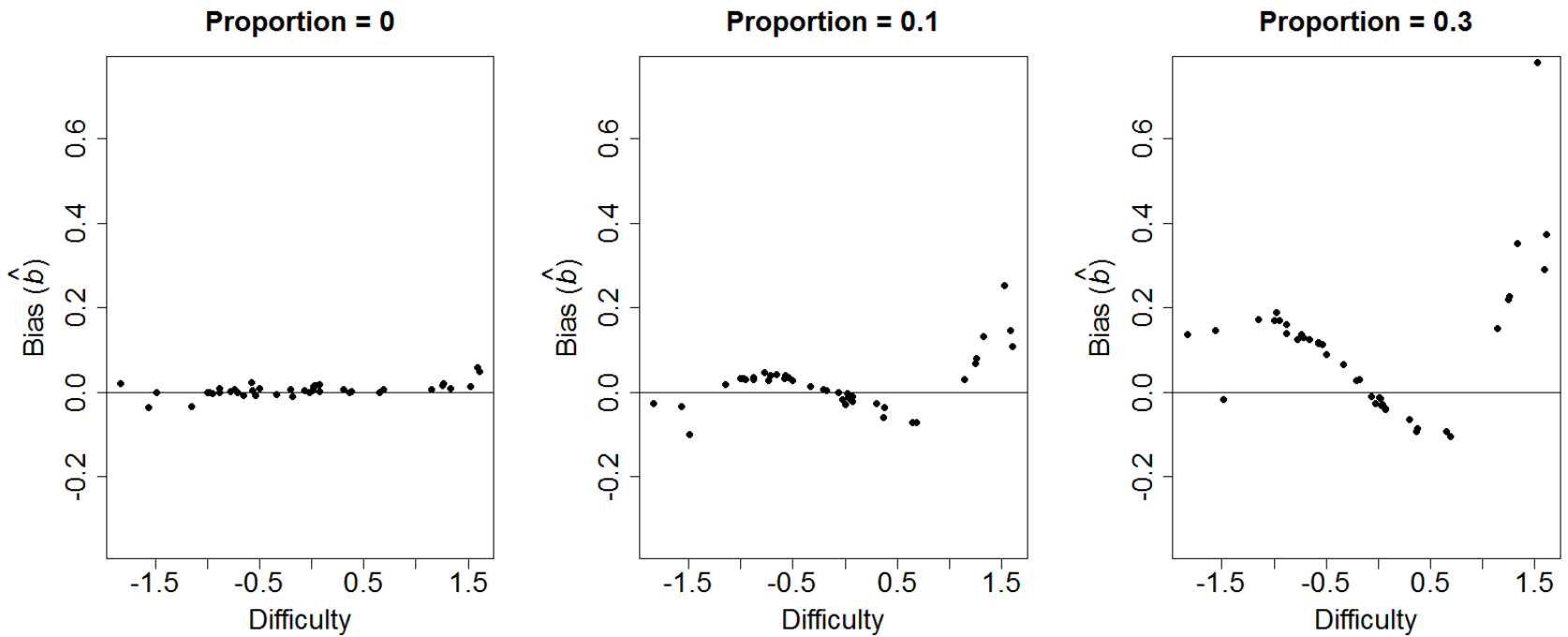
Results: Bias of Discrimination Estimates

Bias of Discrimination Estimates vs. True Discrimination Values



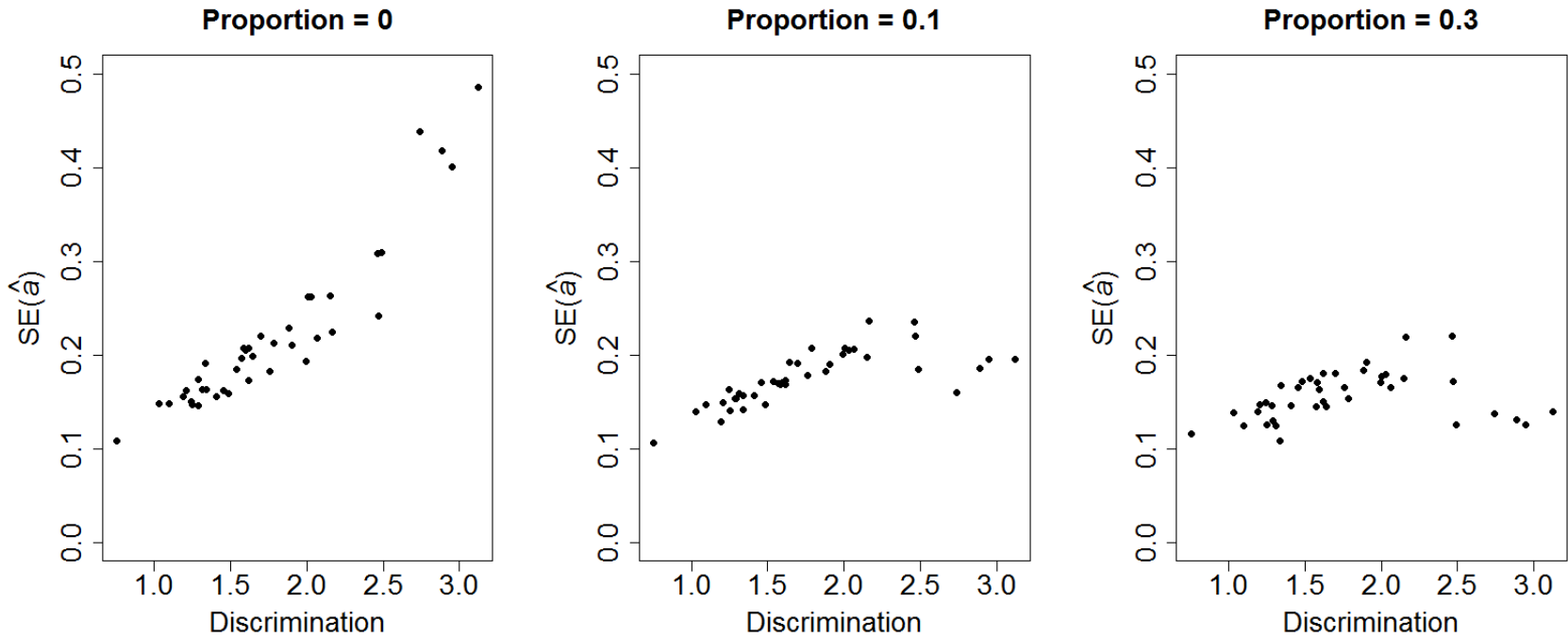
Results: Bias of Difficulty Estimates

Bias of Difficulty Estimates vs. True Difficulty Values



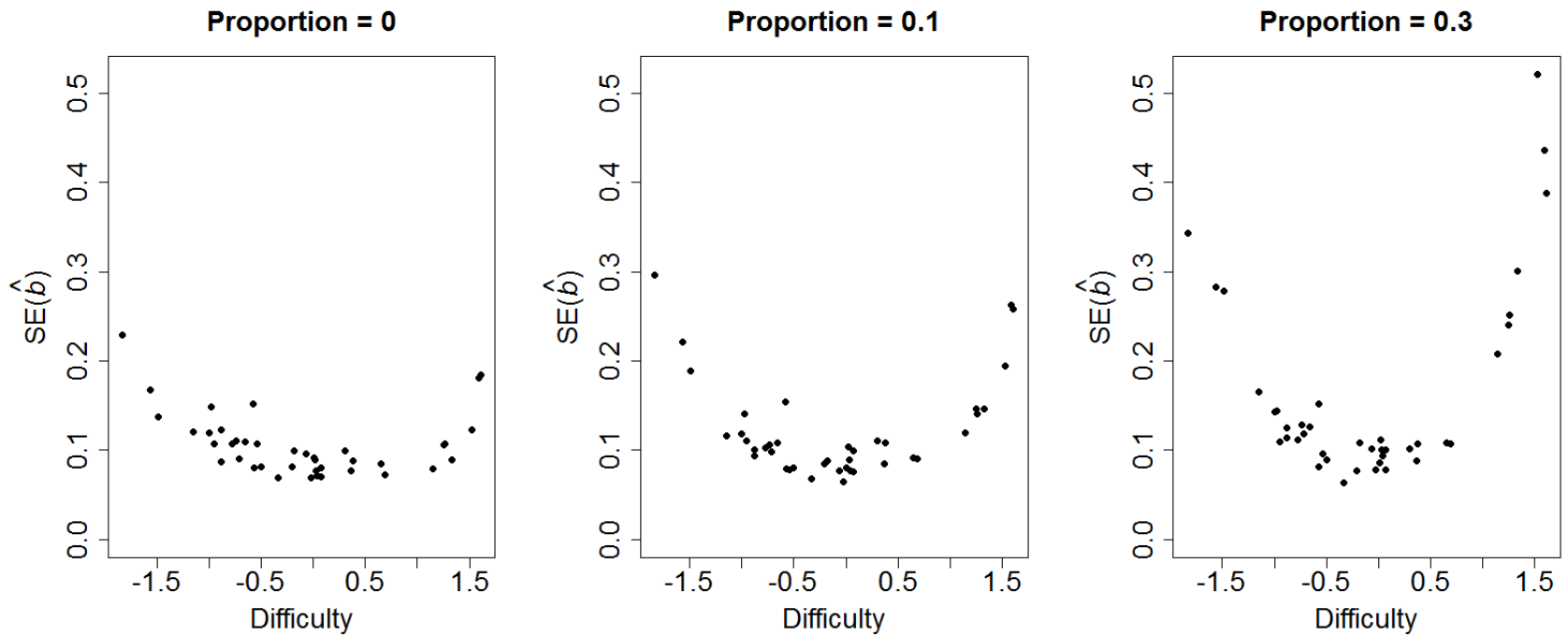
Results: SEs of Discrimination Estimates


SEs of Discrimination Estimates vs. True Discrimination Values



Results: SEs of Difficulty Estimates

SEs of Difficulty Estimates vs. True Difficulty Values



A decorative graphic consisting of a thin yellow circle on the left side. A thick, horizontal olive-green bar extends from the circle across the top of the slide. On the left end of this bar, there is a large black left square bracket. On the right end, there is a large yellow right square bracket.

Study 2

Can we reduce the effects of careless responses by statistically detecting & removing careless responders?

Person-Fit Index: l_z

- standardized log-likelihood person-fit index l_z :

$$l_z = \frac{l_i - E(l_i)}{\sqrt{\text{var}(l_i)}} = \frac{\sum_{j=1}^m w_{ij}(u_{ij} - P_{ij})}{\sqrt{\sum_{j=1}^m w_{ij}^2 P_{ij}(1 - P_{ij})}} \sim N(0,1)$$

- can be re-written so the numerator is a weighted sum of residuals
- Large residuals yield a large negative value for l_z :
 - l_z is used as the test statistic in a one-tailed test
 - if $l_z < -1.65$, examinee exhibits an “aberrant” response pattern

Corrected Person-Fit Index: l_z^*

- l_z is usually evaluated with an examinee's ML ability estimate.
 - but asymptotic $N(0,1)$ distribution only holds when the true ability value is known
 - in practice, variance is less than one → too many Type II errors
- A corrected statistic (l_z^*) follows a $N(0,1)$ distribution when evaluated with an ability estimate:

$$l_z^* = \frac{\sum_{j=1}^m \tilde{w}_{ij} (u_{ij} - P_{ij})}{\sqrt{\sum_{j=1}^m \tilde{w}_{ij}^2 P_{ij} (1 - P_{ij})}} \sim N(0,1)$$

[Sample Cleansing Procedure]

1. Estimate item parameters based on the full, “contaminated” sample.
2. Compute $\hat{\theta}$ and l_z (or l_z^*) for all subjects.
3. Remove subjects with l_z (or l_z^*) < -1.65 .
4. Re-estimate item parameters based on the “cleansed” dataset.

Iterative Cleansing Procedure

1. Use the full sample \mathbf{X}_0 to obtain item parameter estimates $\hat{\gamma}_0$. Use $\hat{\gamma}_0$ to obtain ability estimates $\hat{\theta}_0$ for all examinees.
2. Use $\hat{\gamma}_k$ and $\hat{\theta}_k$ ($k = 0, 1, 2, \dots$) to compute l_z (or l_z^*) for all examinees in the full sample. Create a cleansed sample \mathbf{X}_{k+1} by removing examinees flagged as aberrant.
3. Obtain item parameter estimates $\hat{\gamma}_{k+1}$ based on the cleansed sample. Use $\hat{\gamma}_{k+1}$ to obtain $\hat{\theta}_{k+1}$ for all examinees in the full sample. Substitute $\hat{\gamma}_{k+1}$ and $\hat{\theta}_{k+1}$ into Step 2.
4. Repeat steps 2 & 3 until the proportion of examinees that change classification (i.e., from aberrant to normal, or vice versa) does not exceed .01.

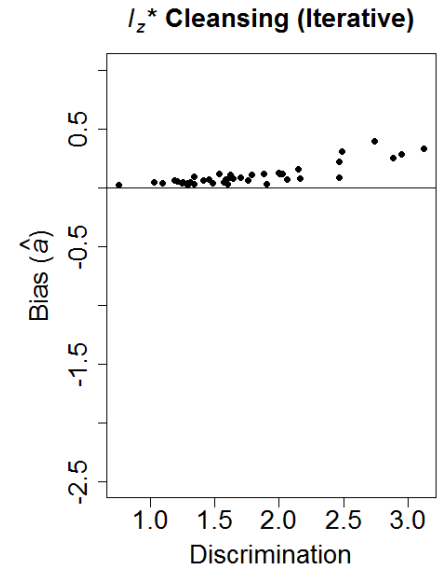
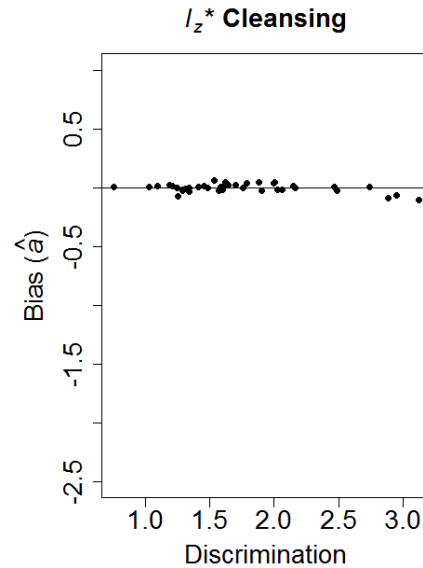
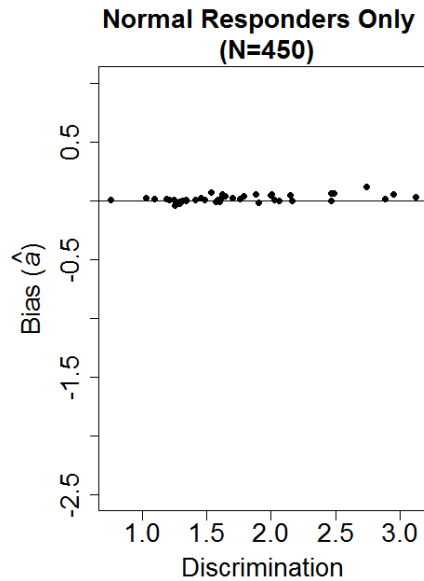
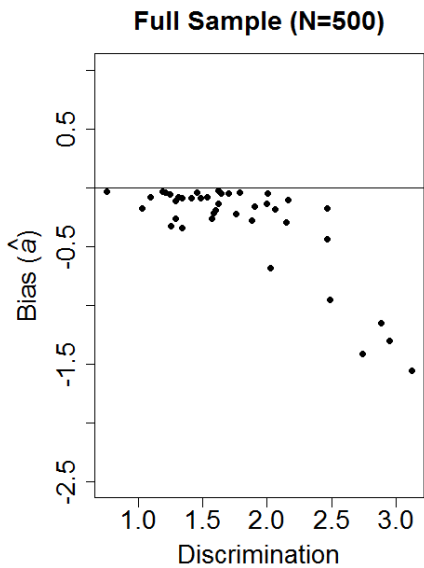
[Iterative Cleansing Procedure]

- Note: if an examinee is removed from the sample in one iteration, they can still be included in the next iteration.
- We repeatedly use “improved” item parameter estimates to compute fit statistics for the full sample.
- Once the examinee classifications (i.e., aberrant or normal) have stabilized, this is evidence that the item parameter estimates have stabilized (which is the ultimate goal).

[Method]

- Use the datasets generated for Study 1
 - $N = 500$, $p = 0\%$, 10%, or 30%
- For each condition, obtain item parameter estimates based on:
 - full, contaminated sample
 - normal responders only (baseline condition)
 - non-iterative and iterative cleansing procedures
 - l_z and l_z^* person-fit statistics ($\alpha = .05$)
- Outcomes:
 - bias & SEs of item parameter estimates
 - Type I & Type II error rates of fit statistics

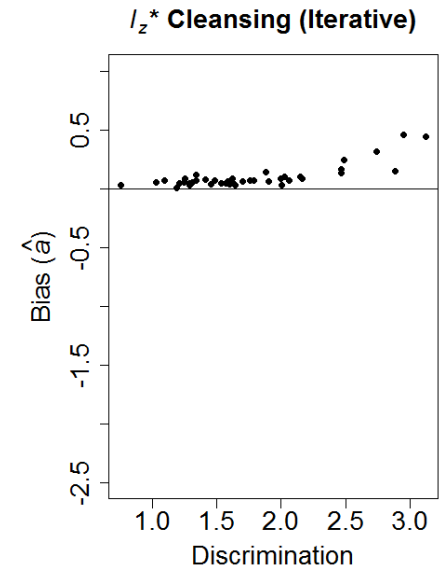
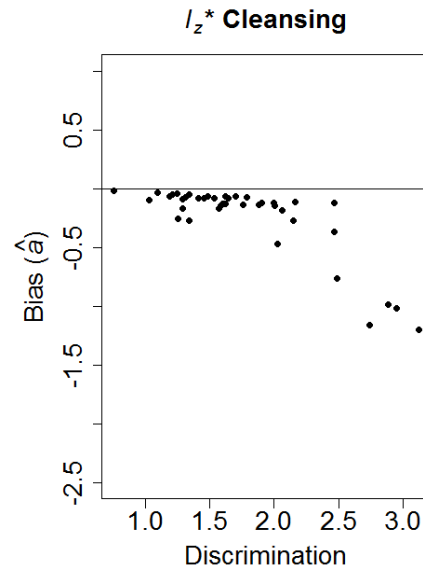
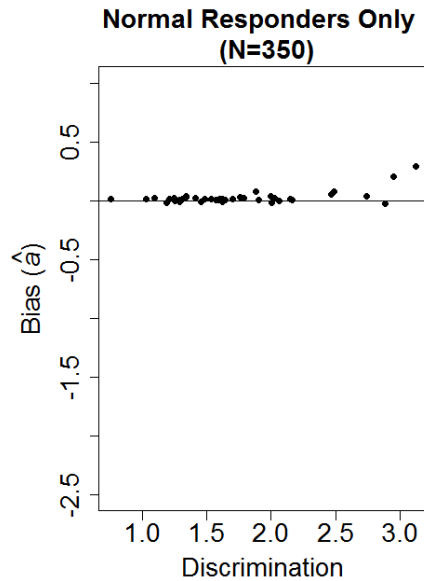
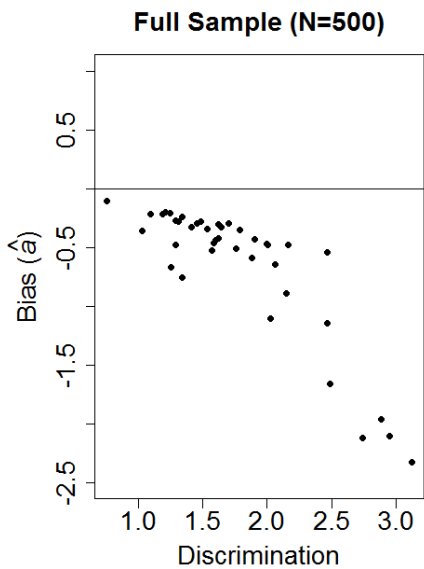
Bias of Discrimination Estimates ($N = 500, p = 0.1$)



Performance of l_Z^* ($N = 500, p = 0.1$)

	Cleansing Procedure	
	Non-iterative	Iterative
# (proportion) flagged	55 (.11)	82 (.16)
P(careless flag)	.88	.61
Type I error rate	.016	.072
Power	.96	.99

Bias of Discrimination Estimates ($N = 500$, $p = 0.3$)



Performance of l_Z^* ($N = 500, p = 0.3$)

	Cleansing Procedure	
	Non-iterative	Iterative
# (proportion) flagged	117 (.23)	173 (.35)
P(careless flag)	.98	.87
Type I error rate	.006	.067
Power	.76	.99

[Conclusions]

- l_Z and l_Z^* perform similarly, regardless of cleansing procedure and the proportion of careless responders
- $p = 0.1$:
 - non-iterative procedure performs better (power = 0.96, Type I error rate close to zero)
 - iterative procedure mainly serves to increase Type I error rate (but is closer to the nominal level)
- $p = 0.3$:
 - iterative procedure performs better (power = 0.99, Type I error rate close to nominal level)
 - non-iterative procedure makes too many Type II errors

[Conclusions]

- Concerning the non-iterative procedure:
 - item parameters estimates used to compute l_z are “contaminated”
 - level of contamination is greater for larger p
 - thus, non-iterative procedure may suffice when p is expected to be low

- But the iterative procedure appears to be the better choice:
 - iterative cleansing produces a relatively “uncontaminated” set of item parameter estimates
 - achieves nominal Type I error rate, regardless of p

[Future Directions]

- Explore effects of careless responding on ability estimates and classification accuracy (i.e., propagation of error).
- We simulated a simple, but extreme type of careless responding.
 - the person-fit statistics performed rather well
 - simulate more realistic careless behavior (not all of an examinee's responses are careless)
- l_Z^* incorporates a correction for not knowing true ability.
 - when item parameters are poorly estimated, may not perform well (using the non-iterative procedure)
 - iterative procedure is promising; an alternative is to derive an additional correction for not knowing true item parameter values