The Role of Identity Item Design in Promoting Survey Data Quality: A Computational Simulation of Gender Identity Questions and Non-Binary Gender Identities

THE UNIVERSITY OF BRITISH COLUMBIA

Kyle Dewsnap
Edward Kroc
Bruno Zumbo

Highlights

- Survey questions can fail to fully capture a person's thoughts, values, and feelings
- Monte Carlo simulations can demonstrate how survey questions about gender identity can lead to bias in regression analysis.
- Further bias appears in scenarios where the categorical variable is correlated with a continuous predictor variables.
- Accurate survey design is crucial to reflect true group differences and effects in regression models.

Summary

Survey research is a popular tool within the social sciences, yet little is known about how measurement error impacts survey data. This study explores the relationship between survey item design and data quality, specifically focussing on gender identity questions for non-binary identities. We designed computational simulations to show how imprecise survey questions can distort estimated differences between majority and minority groups. We specified a scenario where non-binary individuals must choose between less representative "M" or "W" options and show how this can either inflate or attenuate the estimated differences between the groups. Additionally, the precision of these responses is revealed to cause bias for estimates of unrepeated continuous effects, should these variables be correlated with group membership. These findings underscore the importance of precise survey design to accurately reflect true group differences and effects in regression models.

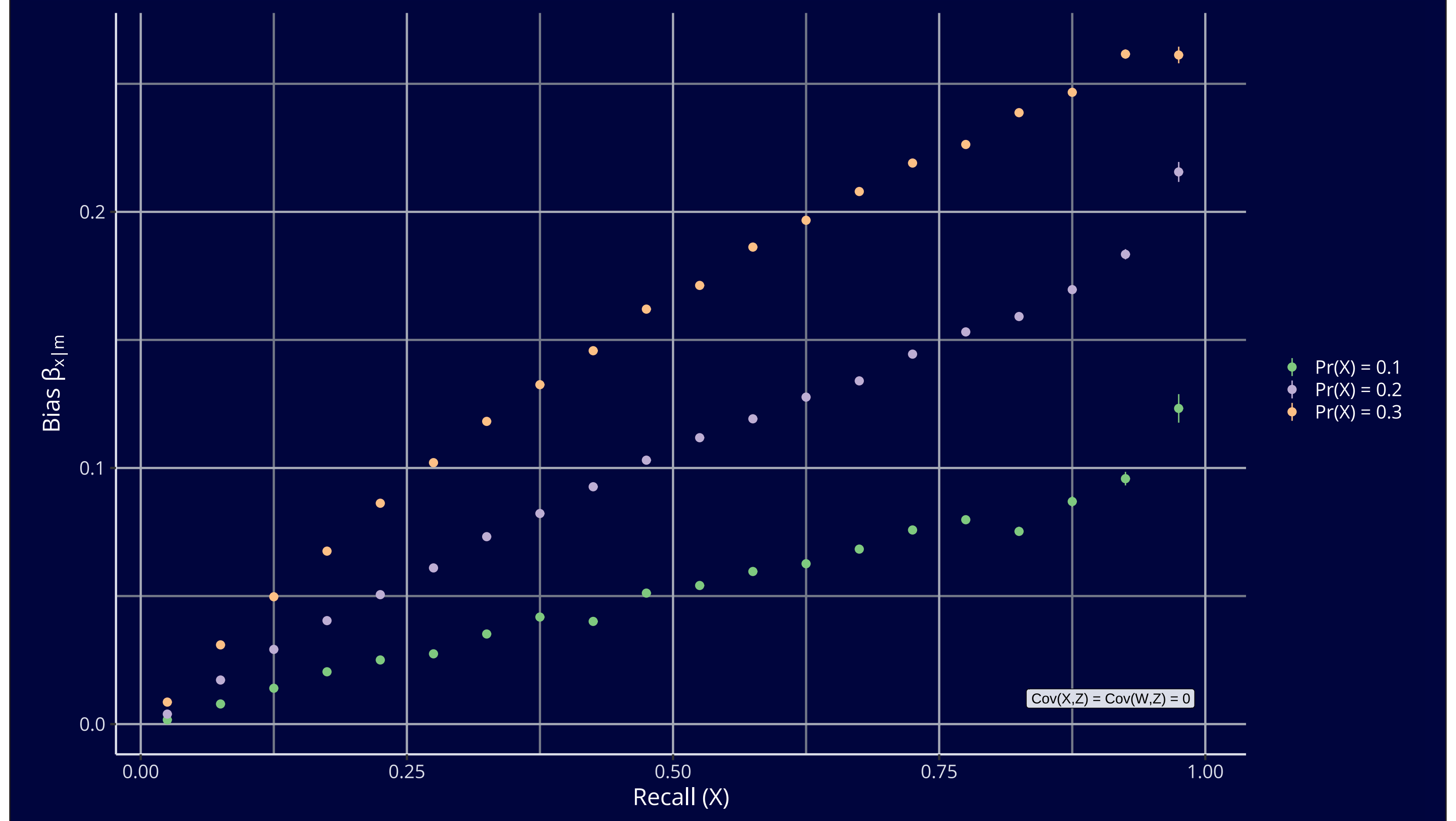


Figure 1: The relationship between recall in the X response option (horizontal axis) and bias in the X or M regression parameter (vertical axis). Different colours represent three levels of proportion of group X members within the sample (0.1, 0.2, and 0.3). Across all three proportion levels bias becomes more positive as recall decreases, but higher proportion levels experience more extreme increases than lower levels.

Miscategorization and Dummy-Encoded Categorical Variables

Imprecise survey questions can inflate the estimated difference between a majority group and a minority group. As we simulate more non-binary people endorsing the M option, we notice a growing, positive bias in the regression coefficient comparing the M and X groups. We also observe cases where all X-participants are hidden in the survey data, making any difference impossible to estimate.

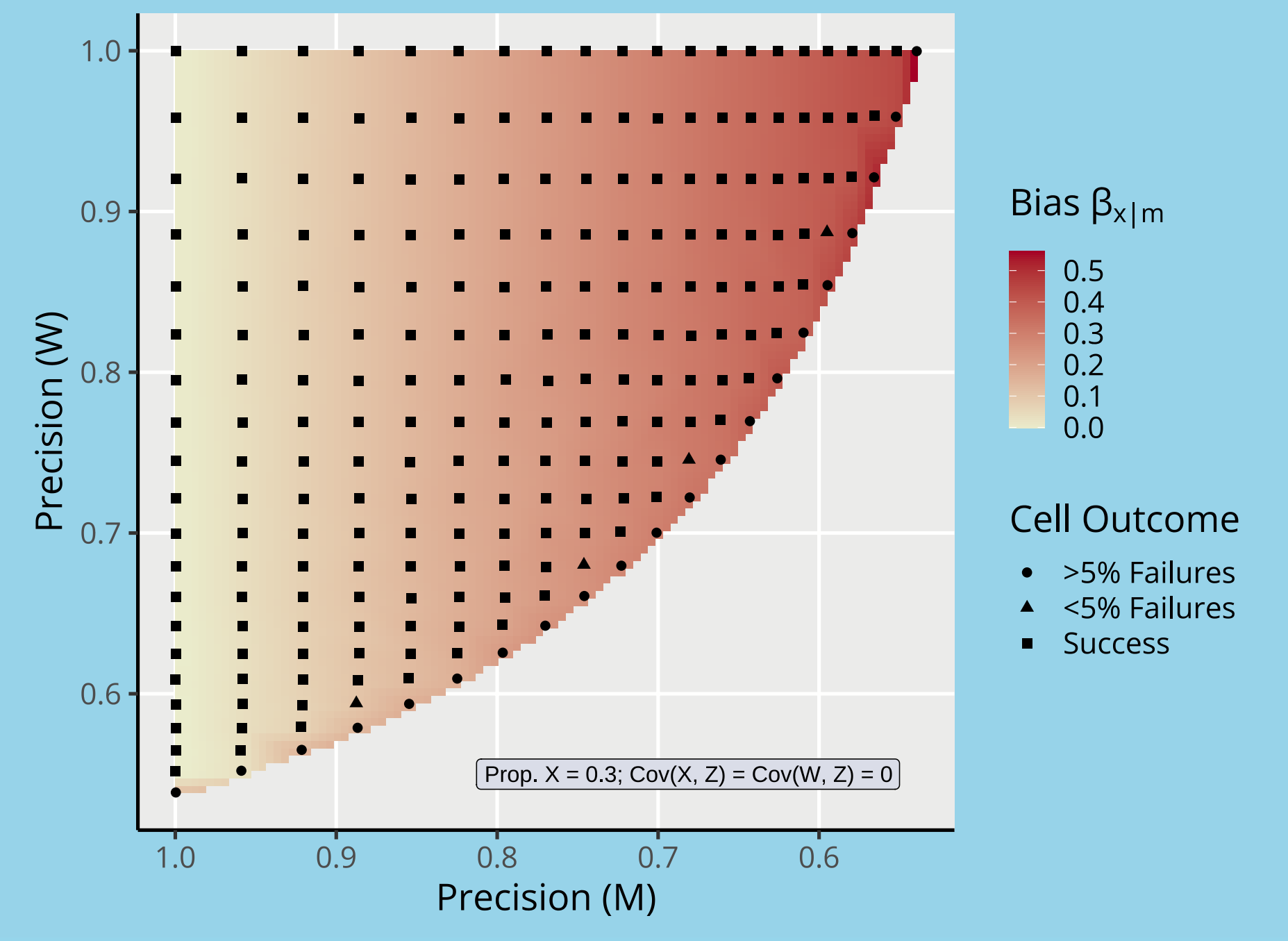


Figure 2: Bias in the X or M regression parameter with points representing each simulation cell. The axes represent the precision of the M and W response options. The colour of the tile represents the level of bias: Bias is shown as becoming more positive as M responses become less precise. The shape of the points shows the number of trials that failed to estimate the regression parameter: Failures are more common with lower precision in both the M and W response options.

Imprecise survey questions can inflate or attenuate the differences between majority groups. As more non-binary people endorse the M option, bias in the coefficient between the M and W groups becomes more positive. We found the inverse happens as more non-binary people endorse W, demonstrating the challenge of using miscategorized data to isolate genuine group differences.

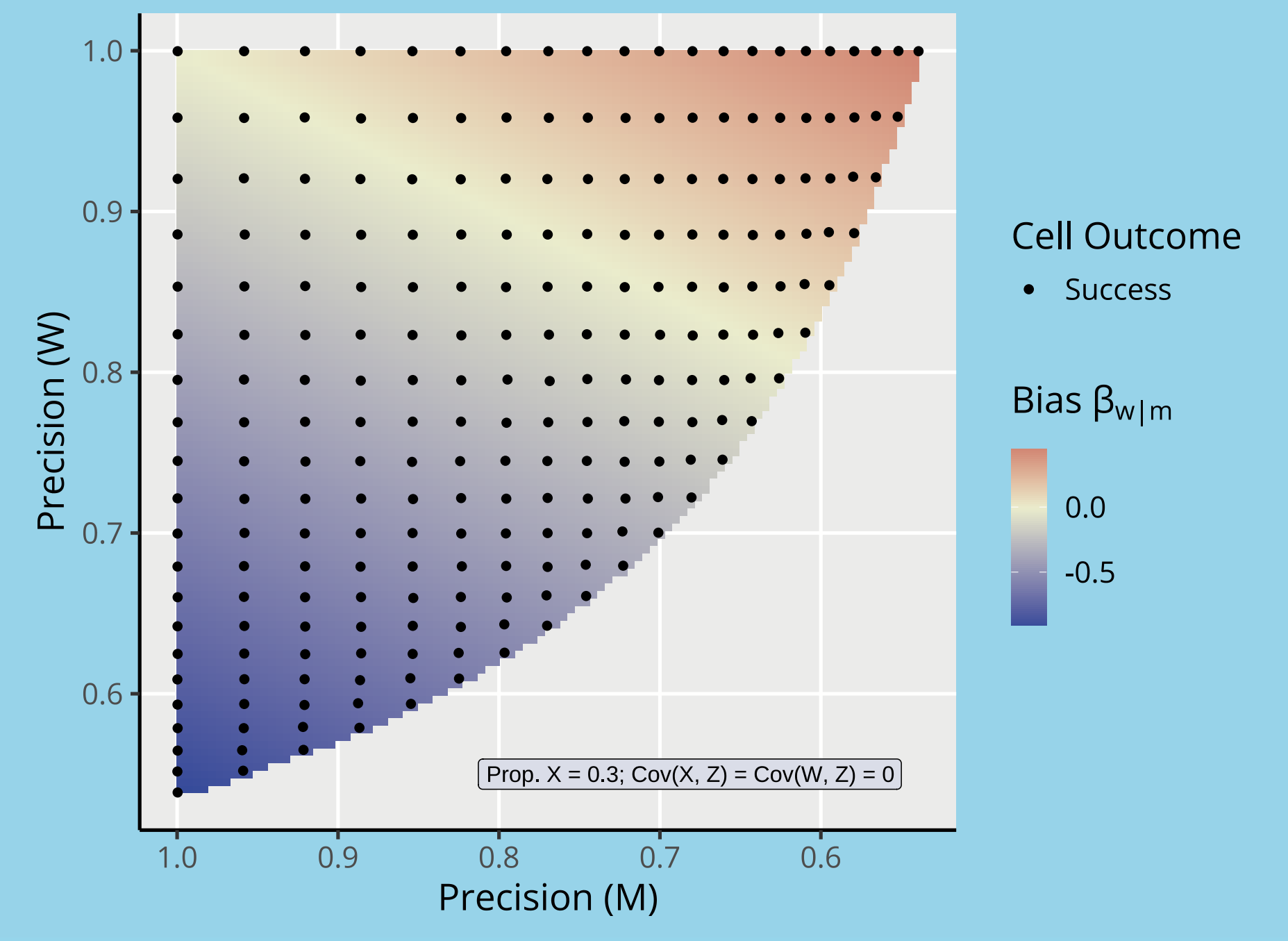


Figure 3: Bias in the W or M regression parameter with points representing each simulation cell. The axes represent the precision of the M and W response options. The colour of the tile represents the level of bias; bias becomes more positive as M responses become less precise, and more negative as W responses become less precise. The points indicate that no trials failed to estimate the regression parameter.

Miscategorization and Continuous Variables

Using imprecise survey responses alongside other variables can further bias a regression model. In several scenarios, we allowed the continuous Z variable to correlate with group membership while always maintaining its mean of zero. As we increased this correlation, we observed how changes in the precision of the survey responses both attenuate and inflate estimates of Z's effect on the response.

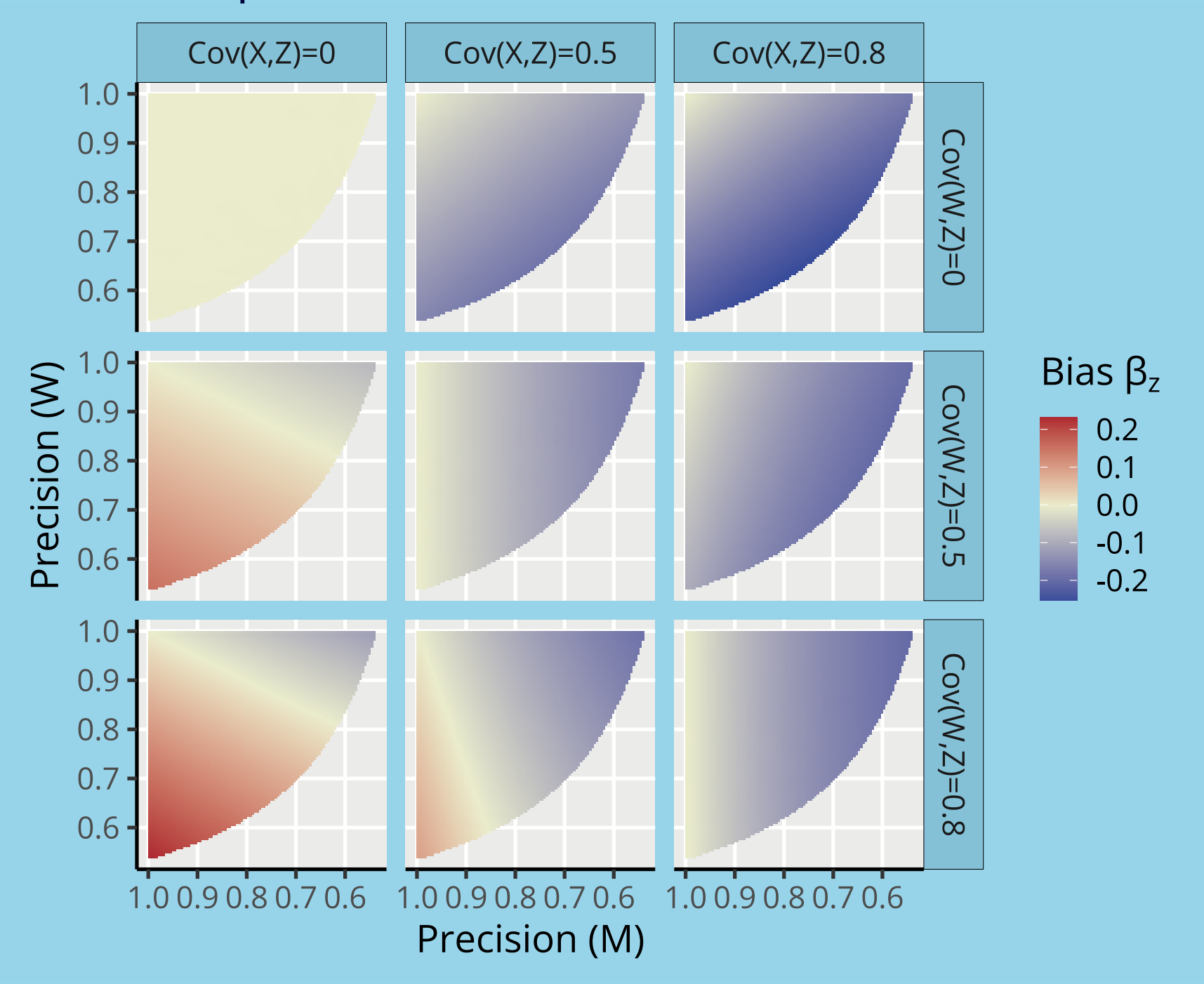


Figure 4: Bias in the Z regression parameter, panelled across three levels of covariance between the Z variable and membership in either the X or W groups. The axes represent the precision of the M and W response options. The colour of the tile represents the level of bias: At no covariance the tiles are uniform and indicate no bias. As covariance in X and W increases, the plots shift to indicate more negative and more positive bias, respectively, across different levels of precision.

Important Concepts

Precision: The proportion of people who responded as belonging to a certain category that truly belong to that category. Precision (M) tells us the probability that a participant genuinely belongs to the M group, given that they selected the M response option.

Recall: The proportion of people who truly belong to a certain category that selected the correct survey response option. Recall (X) tells us the probability of a participant selecting the X response option, given that they truly belong to the X group.

$$M, W, X = \text{Categories}$$

$$M^*, W^*, X^* = \text{Response options}$$

$$Precision(M) = \frac{\text{No. M who endorsed M*}}{\text{Total No. of M* endorsements}} = Pr(M|M^*)$$

Simulated model:

$$Y = \beta_0 + \beta_Z Z + \beta_{M|W} W + \beta_{M|X} X + \epsilon$$

$$\epsilon \sim N(0,1)$$

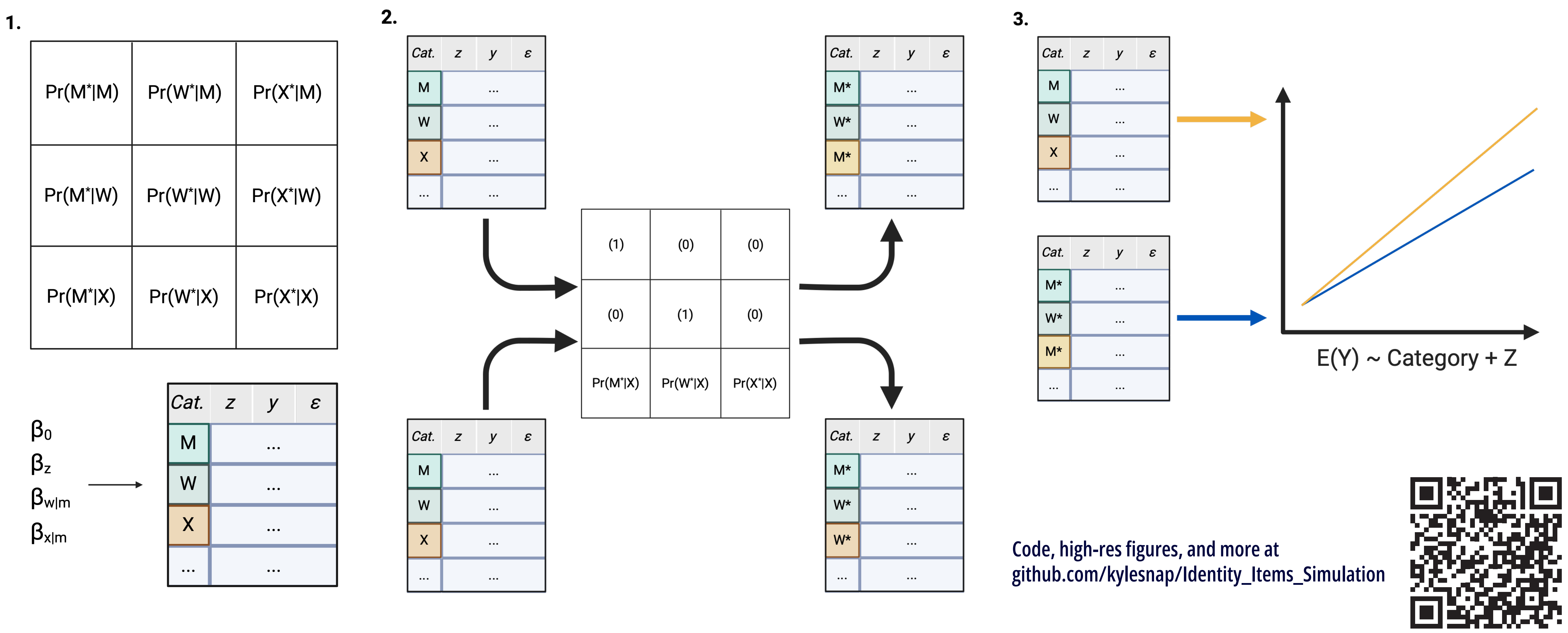
$$\beta_0 = 0, \beta_z = 1, \beta_{M|W} = 1, \beta_{M|X} = -1$$

$$Recall(X) = \frac{\text{No. X who endorsed X*}}{\text{Total No. of X in sample}} = Pr(X^*|X)$$

Additional information: Definitions for precision and recall, along with a description of the simulated (data-generating) model. The data-generating model is set so that Z and membership in W increases expected response (Y) by one, and membership in X decreases expected response by one. Residual errors follow the standard normal distribution. Simulations are performed with 1000 participants, with 1000 replications.

Methods

1. **We created simulated data according to population-level parameters and a regression model.** To describe the survey response process, we used a transition matrix that contains the probability of being categorized into options M, W, or X, conditional on a participant's actual category membership.
2. **The simulated data is transformed by the transition matrix, representing participants answering a gender identity item.** In this scenario, only participants in group X (the non-binary category) can endorse a non-X response option.
3. **We compare the original and transformed data and fit them to identically specified regression models.** By observing how bias between the two models emerges across different transition matrices and population parameters, we infer how miscategorization can alter the results of regression analysis.



Code, high-res figures, and more at github.com/kylesnap/Identity_Items_Simulation