Reducing Non-Response Bias with Survey Reweighting: Applications for Online Learning Researchers

René F. Kizilcec
Department of Communication
Stanford University
kizilcec@stanford.edu

ABSTRACT
In many online courses, information about learners is collected via surveys for accounting, instructional design, and research purposes. Aggregate information from such surveys is frequently reported in news articles and research papers, among other publications. While some authors acknowledge the potential bias due to non-response in course surveys, there are no investigations on the severity of the bias and methods for bias reduction in the online education context. A regression-based response-propensity model is described and applied to reweight a course survey, and discrepancies between adjusted and unadjusted outcome distributions are provided.

Author Keywords
Survey research; Survey reweighting; Non-response bias

ACM Classification Keywords
H.5.2 User interfaces: Evaluation and Methodology

INTRODUCTION
Recent online courses have seen massive enrollments from around the world by people of various backgrounds. There has been an interest from course instructors, researchers, and the wider public to find out who takes these courses and for what reasons. Moreover, online learning researchers are keen to ask online learners questions about their experience and collect other self-report measures. As a result, online learners are frequently asked to complete one or more surveys before, during, and after a course.

These surveys tend to be optional to comply with policies of ethical review boards and to remain consistent with the open nature of many online courses. A natural consequence of optional surveys are self-selection effects, as people decide whether or not to respond to the survey and their decisions are rarely random (i.e. a fair coin flip). Instead, self-selection can lead to non-response bias, when those who respond to the survey differ in the outcome variable from those who do not respond. In other words, the decision to respond is not independent of the survey measures. For instance, if satisfied online learners are more likely to respond to the survey than dissatisfied ones, then we would expect an upward bias in survey questions on learner satisfaction.

The consequences of non-response bias can be detrimental for the validity of research that relies on survey measures. Fortunately, there exist statistical methods that address the issue by reducing the amount of bias. In the context of online course surveys, where demographic information at the population level (e.g., Census data) is usually unavailable, many common survey adjustment techniques cannot be applied. However, there is a promising alternative method that utilizes learners’ behavioral data, which is available at the population level, to weight responses according to respondents’ likelihood to respond. In the following sections, I elaborate on this method of survey reweighting and provide an exemplary application.

SURVEY WEIGHTS
There is a large literature on survey non-response which proposes different adjustment methods depending on the available data and research goal (see [2, 4] for reviews). Gelman and Carlin [3] provide further details and a discussion of propensity reweighting, the type of technique used here, and poststratification, an alternative approach. Yet another alternative technique is multiple imputation as advocated by Rubin [5].

The adjustment technique that is described and tested here employs a response-propensity model based on penalized logistic regression. The response-propensity, the likelihood of responding to the survey, is used to weight survey responses, such that those with a low propensity receive higher weights and vice versa when outcome distributions and specific statistics are computed. The propensities are computed using a logistic regression on the learner population where survey non-response is the binary outcome and various behavioral variables are predictors. To eliminate noise from predictors with low predictive power or that are highly correlated, a penalized regression is used to perform the variable selection task. In the following application, response propensities are estimated with the elastic net penalty [6], using a the weighted combination of the ridge and lasso penalty that minimizes the cross-validated error.
APPLICATION
In a ten-week massive open online course (MOOC) on a topic in Sociology with 53,077 enrolled learners, a course survey was announced at the end of the first week followed by regular reminders until the third week. 9,583 learners responded to the survey (18.1% response rate) which contained several demographic and course specific questions. The behavioral variables that were used as predictors for computing the response-propensities were

- **video logs** number of distinct lectures viewed and play, pause, and seek events
- **page view logs** number of views for each course page
- **forum logs** number of posts, comments, thread views, up-votes, and downvotes
- **basic account data** registration and last access time, an indicator for each timezone, and whether the learner unenrolled.

The model was fit and cross-validated using the \texttt{cv.glmnet} function from the \texttt{glmnet} R package \cite{Friedman2010} and weights were computed by taking the inverse of the estimated response probabilities from the \texttt{predict.glmnet} function with arguments \texttt{type=“response”} and \texttt{s=“lambda.min”}.

RESULTS & DISCUSSION

Variable Selection
Only a small number of mostly lecture-related variables was predictive of survey response. The following seven out of 292 predictors were selected by the elastic net: last course access time; number of distinct lecture views; video pause events; and views of the course landing page, the lecture browsing page, the lecture player page, and the course wiki page. Surprisingly, no geographic or forum-related predictors were selected.

Model Fit
For a simple evaluation of the model fit the predicted response outcome is compared with actual response behavior (without cross-validation). 28.3% (98.0%) of those who did (not) take the survey were correctly predicted as (non-)respondents, yielding a correct prediction in 86.9% of cases. Thus, the model might overfitted towards correctly predicting non-response.

Adjusted Outcome Distributions
The effect of survey reweighting on the age distribution is illustrated in Figure 1, which suggests that learners under 40 were generally under-represented in the survey, while those over 40 were over-represented.

Contrary to the frequently observed pattern of higher survey response rates among women, the proportion of female learners was adjusted downwards by 1.25% points, while that of males increased by 0.32% points, i.e., males were more likely to respond than females.

Finally, on the question about learners’ highest level of formal education, the proportion of those with undergraduate degrees increased (+1.63% points), while the proportions of those with postgraduate/graduate degrees (~2.17% points) and Ph.D.s (~0.43% points) decreased. Adjustments for other categories were only minor (~0.2% points).

CONCLUSION
The application of survey reweighting to a typical course survey produced shifts in three key demographic features. Therefore, online education researchers who use survey data should address non-response bias in online course surveys to ensure that the conclusions they draw are valid. The reweighting method presented here is a first step towards an adequate solution for addressing this issue. These findings require further validation with other surveys and adjustment methods, e.g., multiple imputation \cite{Rubin1987}.

ACKNOWLEDGEMENTS
I am grateful to Dean Eckles for his advice on survey adjustment methods, Emily Schneider for general feedback, and the Office of the Vice Provost for Online Learning for supporting this research.

REFERENCES