Proceedings of the 89th Annual International Meeting of the Psychometric Society, Prague, Czech Republic (2024).

Accounting for ceiling effects in gender equality endorsement. A zero inflated modeling approach.

D. Carrasco,[*1] D. Torres Irribarra[2], N. López Hornickel[3], and A. Sandoval-Hernández[3]

[1] Facultad de Educación, Pontificia Universidad Católica de Chile, Santiago, Chile

[2] Escuela de Psicología, Pontificia Universidad Católica de Chile, Santiago, Chile

[3] Department of Education, University of Bath, Bath, United Kingdom

*Corresponding author. Email: dacarras@uc.cl

Abstract

Gender equality endorsement is an intergroup measure present in various survey-based studies and is a prominent indicator among the Sustainable Developmental Goals (SDG) (Sandoval-Hernández et al., 2020). To this end, countries can rely on the gender equality endorsement scale included in the International Civic and Citizenship Study (ICCS) (Schulz et al., 2018), which provides probabilistic samples of 8th-grade students from different countries and assesses gender equality endorsement between men and women. Traditional methods for generating scores with this scale rely on the partial credit model (PCM), a response model that utilizes a normally distributed latent variable to represent students' propensity to respond to the various items included in the instrument. Moreover, researchers rely on regression models to address research questions about related factors and the effects of program evaluation. However, the scale scores of this instrument are highly skewed. This skewness is desirable. It means a noticeable portion of students endorse gender equality at the scale ceiling. Nevertheless, traditional regression models may produce distorted estimates in the presence of ceiling effects on the total scores. We propose a method that relies on the monotonicity property of the PCM scores and a reverse sum score. We use zero-inflated models to separate ceiling cases from the rest of the scores, allowing us to make inferences on both sides: the students at the ceiling and those in the remainder of the distribution. This method is a helpful tool for program evaluations dealing with ceiling effects in their attribute of interest.

Keywords: gender equality, sustainable developmental goals, large-scale assessment

1. Introduction

Different response variables in social science may exhibit high skewness. Student bullying and aggression victimization may present low prevalence in such a way that if a random sample is drawn from the population of students, a good proportion of students may not report suffering from aggressive events at school, e.g., 45% of students report not suffering from severe aggression events at school, across six different countries from Latin America (Carrasco, Torres Irribarra, et al., 2023). In such a case, the left side of the scale score distribution would accumulate more cases than the proportion of cases on the right side of the mean. Other skewed variables may exhibit censoring when the instrument is unable to capture the range of the attribute. Suppose an instrument was designed to screen depressive symptoms on clinical samples (e.g., "I did not attend to work because I did not have the energy"). In this case, the instrument may produce excessive zeros in nonclinical samples because non-depressive participants may not respond positively to any item (Boulton et al., 2018). The same is possible at the other end of the scale. Participants may express high agreement with all the items of an intergroup attitudinal scale battery, thus indicating the maximum endorsement of equality. For example, between countries, the mean estimate of students who are more likely to agree to all items expressing gender equality across 23 countries is 48% (López-Hornickel et al., 2024). Consequently, the distribution of such scores is highly skewed, with cases accumulating at the ceiling.

These different examples can be viewed as different cases of censored variables. These are variables that appear to be capped at a certain threshold, unable to display the true value of the attribute of interest in the observed score. The depression example is a case of left censoring, where the true value of the attribute may lay before the observed minimum of the instrument score. In contrast, the intergroup attitude scale score example is a case of right censoring, where the instrument-observed scores display an upper bound, lumping cases at the observed maximum. Finally, the zeros in the bullying example can be interpreted as the true standing of the attribute, as the absence of a bullying event. This latter scenario is referred to as a semicontinuous variable, where zero is considered a true zero rather than the result of scale score truncation (Boulton et al., 2018).

Censored variables and semicontinuous variables pose challenges for linear regression models. Linear regression models assume that residuals are normally distributed and homoscedastic. As such, the residuals around the expected values generated with the fitted model will display common variance across the values of the predictor. However, the generated values with the fitted model can differ significantly from the observed values when the response variable is highly skewed, especially in cases with ceiling or floor effects, where a substantial portion of cases accumulate at the maximum (ceiling) or the minimum (floor) score. Finally, highly skewed variables often present outlying observations with high leverage, that is, observed extreme values at the other extreme of the censored side. As a consequence of the accumulation of cases in one extreme and the presence of outlying values with high leverage, regression estimates can result in bias (McBee, 2010). This bias can be so substantial as to render the estimation of intervention effects close to a null (Boulton et al., 2018). In summary, regression models are ill-equipped to model highly skewed response variables.

Besides normal theory regression models, other alternatives could be applicable. One could recode the response variable to separate ceiling cases from the remainder and, on the newly binary variable, use a logistic regression. However, such an approach would lose information and would also result in biased slopes. Another alternative is Tobit regression, which can get corrected slopes despite the censoring of the response variable (McBee, 2010). Tobit regression assumes a latent variable where participants would have the non-observed true value if there were no censoring. It assumes that the regressor coefficients are the same for the latent variable and the observed variable while accounting for censoring (Wilson et al., 2020). However, this model is not informative for the cases at the extremes. The model can produce regressor effects as if there were no censoring. However, it does not provide estimands for the relationship between regressors and the proportion of values at the extreme of the scale because it assumes a single data-generating mechanism. In the current study, we propose fitting a zero-inflated negative binomial model to account for ceiling effects (Loeys et al., 2012) in the scores on students' endorsement of gender equality (Schulz et al., 2018), a highly skewed scale score. In this application, we propose using the total score of the scale as a reverse sum score. The cases at the ceiling will be represented as zeros, the case at the highest endorsement of gender equality. Simultaneously, the non-ceiling cases will be represented as the remainder of the sum score, which we interpret as indicating that higher sum scores correspond to greater endorsement of sexism. As such, the chosen model can provide estimates of the relationship between the covariate and cases that reach the highest score or not, as well as those with higher sexism endorsement, given that these cases do not reach the ceiling.

Students' endorsement of gender equality (Schulz et al., 2018) is a highly skewed variable characterized by a high accumulation of cases at the ceiling. These scores are realizations generated with a fitted partial credit model, scaled to a mean of 50 and a standard deviation of 10. Participants with the highest scores have an average of 63.94 on this scale. In the pooled sample, 31% of the participants reach the observed maximum. The proportion of cases at the ceiling varies between countries from 4% to 60%. Figure 1 illustrates these ceiling cases for four countries out of 24 samples, highlighting in black the observations at the maximum score. We are including exemplary countries with low, medium, and high proportions of cases at the ceiling.

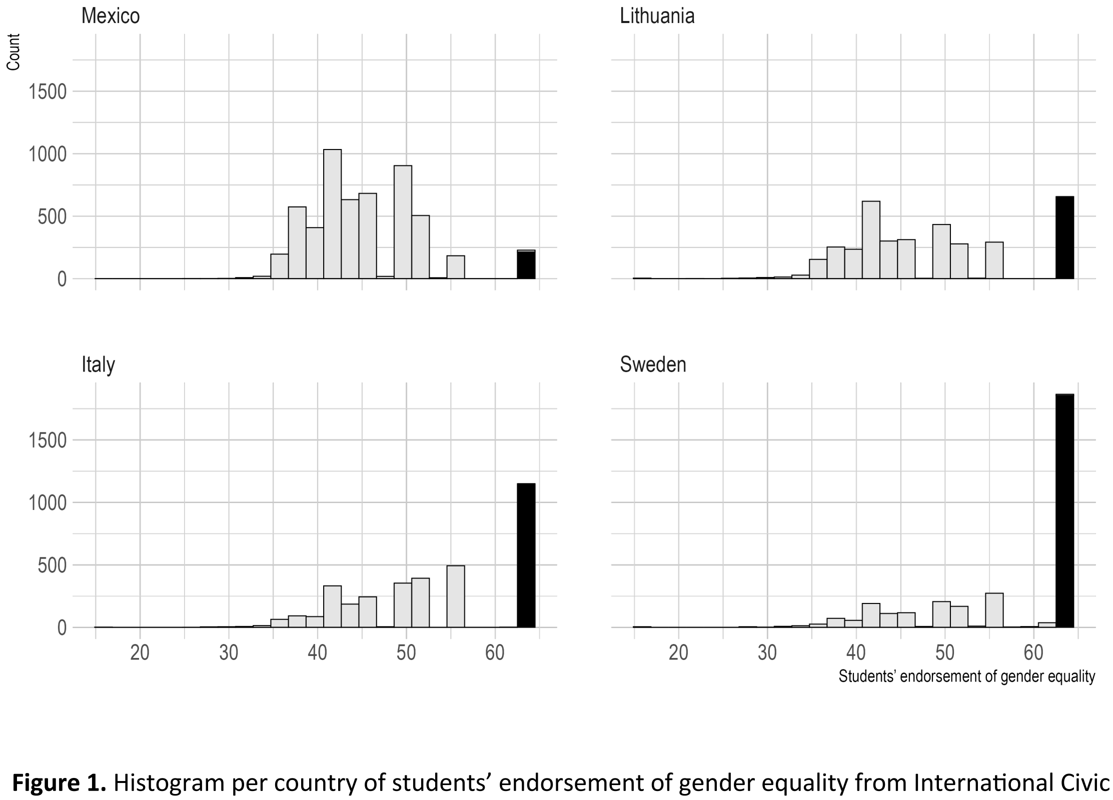


Figure 1. Histogram per country of students' endorsement of gender equality from International Civic and Citizenship Study 2016

The remainder of the paper is organized into the following sections. In the Ceiling Effects and Censoring section, we discuss two common processes that may underlie the accumulation of ceiling effects, considering the students' gender equality endorsement instrument. Then, in the Modeling Strategy section, we described the proposed method and its rationale. In the Methods section, we describe the data selected for the present illustration. In the Results section, we present the obtained results. Finally, in the Conclusion and Discussion section, we discuss the applicability of zero-inflated models not only to naturally occurring zeros but also to scale scores subject to ceiling effects.

2. Ceiling effects and censoring

Likert-type scales, particularly those addressing endorsement of egalitarian attitudes, such as the gender equality endorsement scale (Schulz et al., 2018), are prone to ceiling effects. That is a high accumulation of observations at the maximum score. One interpretation of this phenomenon is to assume that the trait being measured is not well matched to the scale as if the instrument lacks items or questions capable of registering answers that represent higher levels of the trait of interest. Hence, participants with higher levels of the trait tend to be grouped at the ceiling.

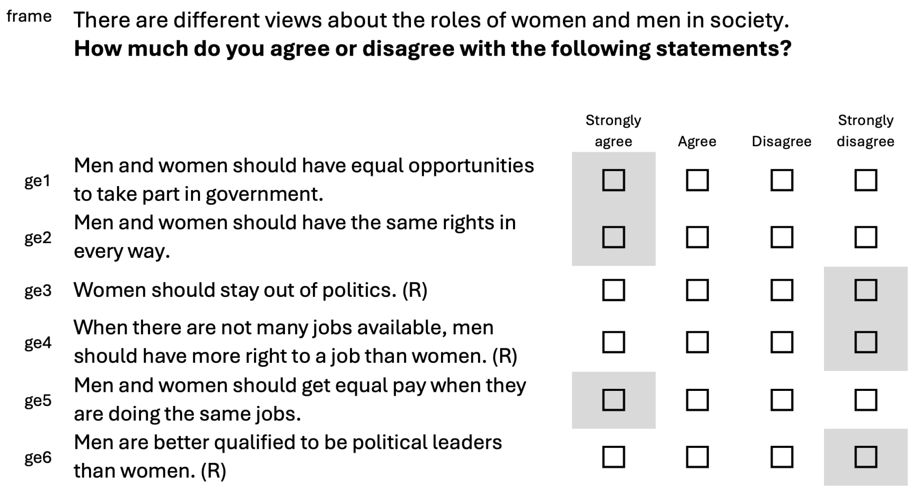
A different explanation is applicable for naturally occurring zeros, where floor effects are observed due to the true absence of the attribute. Measures of substance consumption, such as cigarettes or alcohol, are examples of this scenario, which allow participants to respond indicating zero consumption. In this scenario, non-consuming participants can accumulate as zeroes on the floor of the scale (e.g., Boulton et al., 2018). Such an explanation does not seem applicable to multi-item instrument scores assessing attitudes with polytomous items, such as those measuring endorsement of gender equality (Schulz et al., 2018). That is, to consider cases at the ceiling of the score as cases at the true standing in the attribute of interest. Although item response theory (IRT) models, such as the partial credit model, can generate scores that differ from the sum score of responses, the IRT-generated scores will be censored, resulting in a high accumulation of cases at the right end of the scale.

Regardless of the assumed explanation for the floor or ceiling effects, the accumulation of cases at the ceiling (or floor) needs to be dealt with, especially when covariate effects are of interest.

3. Modeling strategy

Student's endorsement of gender equality is a multi-item scale (Schulz et al., 2018). It consists of three positive and three negative assertions (reverse-coded), to which students indicate their level of agreement using four ordered choices. Scores are generated with a partial credit model (Schulz et al., 2018). In Figure 2, we highlight the highest response category in grey.

Likert-type scales do not produce naturally occurring zeros. However, scores produced with Likert-type scales can present ceiling effects. The partial credit scores are highly skewed and present ceiling effects. In Table 1, we present the pooled sample descriptives of this score. Nearly 31% of the observed cases are at the extreme of the distribution in the pooled sample data of the International Civic and Citizenship Study 2016.



frame There are different views about the roles of women and men in society.
How much do you agree or disagree with the following statements?

		Strongly agree	Agree	Disagree	Strongly disagree
ge1	Men and women should have equal opportunities to take part in government.	□	□	□	□
ge2	Men and women should have the same rights in every way.	□	□	□	□
ge3	Women should stay out of politics. (R)	□	□	□	□
ge4	When there are not many jobs available, men should have more right to a job than women. (R)	□	□	□	□
ge5	Men and women should get equal pay when they are doing the same jobs.	□	□	□	□
ge6	Men are better qualified to be political leaders than women. (R)	□	□	□	□

Figure 2. Students' endorsement of gender equality items highlighting the ceiling pattern of response

In the present study, we propose to leverage the monotonicity property of the sum score and the partial credit score realizations (Kang et al., 2018). These two scores follow a one-to-one relationship. If we recode responses from 0 to 3, from the lowest agreement to the highest agreement for items ge1, ge2, and ge5 (and its reverse pattern for items ge3, ge4, and ge6), the ceiling in the sum score would be 18 points whereas, in the item response theory score of the study, these are 63.94 points. Moreover, if we reverse the sum scores, the ceiling cases would be at zero, relocating the ceiling cases to the bottom of the scale. In Figure 3, we illustrate the one-to-one mapping between the proposed reverse sum score and the original partial credit scores in a scatter-plot with additional histograms in their margins. The ceiling cases from the partial credit score are located at zero in the reverse sum score.

Table 1. Students' endorsement of gender equality pooled sample descriptives

N	Obs.	Mean	SD	Min	P25	P50	P75	Max
94603	0.98	51.31	9.81	16.32	42.64	48.72	63.94	63.94

Notes. N = total number of observations; Obs. = percentage of observed cases; Mean = sample mean; SD = standard deviation; Min = minimum score; P25, P50 and P75 = percentiles; Max = maximum score.

The zero-inflated negative binomial model has two parts. The logit part of the model can help us to address research questions related to students reaching the highest observed level of the attribute given the present instrument. These students are helping to fulfill the Sustainable Development Goal of gender equality (Sandoval-Hernández et al., 2020). While the second part of the model can be interpreted as a measure of "sexism", the remainder can be seen as a measure of the students who do not reach the gender equality endorsement. The higher the value, the more responses denote less gender equality endorsement.

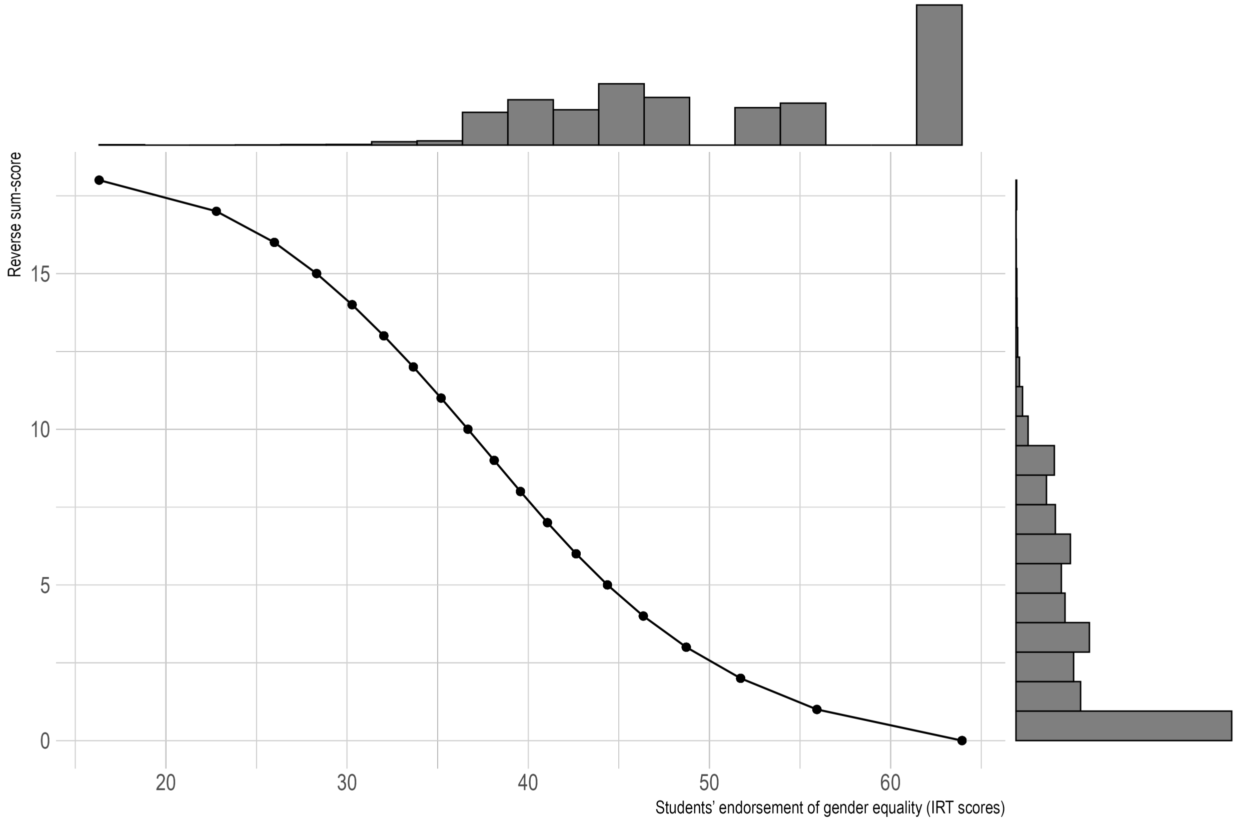


Figure 3. One-to-one relationship between the students' endorsement of gender equality (IRT scores) and the reverse sum score of responses

4. Methods

We use the student sample from Italy in the International Civic and Citizenship Education Study (ICCS 2016) (Schulz et al., 2018). This study employs a two-stage probabilistic design comprising 3,450 observations nested within 170 different schools.

As a response variable, we created a reverse sum score of the student's responses to the scale of endorsement of gender equality, where higher values represent higher sexism, and the zero value relocates the cases to the highest endorsement of gender equality. We first recoded responses to items ge1, ge2, and ge5 as follows: "Strongly disagree" as 0, "Disagree" as 1, "Agree" as 2, and "Strongly agree" as 3. Then, we proceeded accordingly to the reverse items ge3, ge4 a, and ge6, where we recoded responses as follows: "Strongly disagree" as 3, "Disagree" as 2, "Agree" as 1, and "Strongly agree" as 0. Then, we sum all the responses. Finally, we reverse this sum score using the formula: maximum observed value plus minimum observed value minus the observed sum score (18 + 0 - x). Thus, the minimum of this score, the zeros, is the maximum of the partial credit scores and the maximum of the sum score, while a maximum of this scale is the minimum of the sum score and the minimum of the partial credit scores. Figure 3 illustrates the one-to-one mapping between the partial credit scores and the proposed reverse sum score.

For illustrative purposes, we include students' sex (1 = female, 0 = male) as a covariate, mothers' highest level of education (1 = tertiary, 0 = non-tertiary), and a school teaching practice variable called "Open classroom discussion." This school practice is built with students' responses, where students act as informants. Higher scores indicate classrooms where it is more frequent that students are exposed to discussions involving controversial issues, where students are encouraged to express their opinions, and where different sides of the controversies are explained (Carrasco et al., 2018). Scores of this later covariate were obtained using a PCM, which were then standardized at the national level and partitioned into within-school students' deviations from school means and between-school scores' deviations from the grand mean.

We fit a zero-inflated negative binomial model, accounting for the sampling design using Pseudo Maximum Likelihood and Taylor Series Linearization (Asparouhov, 2005; Stapleton, 2008). We are including the schools as primary sampling units and jackknife zones as stratification variables. We chose the zero-inflated negative binomial (ZINB) over a zero-inflated Poisson due to overdispersion of the response variable. The sum score presented a very long tail and a variance estimate exceeding the size of the mean estimate. While the zero-inflated Poisson model is suitable for count and skewed variables with floor effects, it forces the mean and variance to be equal (Muthén et al., 2016). In contrast, a zero-inflated negative binomial is a more flexible distribution that allows the variance to be larger than the mean of the response variable (Green, 2021). To fit the proposed model while accounting for the study's complex sample design, we used Mplus 8.11 (Muthén et al., 2017). We are sharing our code to reproduce the estimates presented in Table 2 (https://doi.org/10.6084/m9.figshare.29323067).

Following Muthén et al. (2016) notation, the ZINB model can be considered a mixture of two classes. In this application, let π_{ij} denote the unobserved probability of participants reaching and exceeding the maximum score of gender equality endorsement (excess zeroes on the sexist responses). We are indexing observations under “ij” for students (“i”) within schools (“j”). Complementary, $1 - \pi_{ij}$ is the probability of being in the class that follows a negative binomial for the rate of sexists’ responses, where zero counts and positive counts can be expected. Thus, the probability of all zeroes follows expression (1), where $e^{-u_{ij}}$ is generated with a negative binomial.

$$\Pr\left(y_{ij} = 0\right) = \pi_{ij} + (1 - \pi_{ij})e^{-u_{ij}} \tag{1}$$

The ZINB logit part of the model conditions π_{ij}, the excess zeroes representing the ceiling cases of our response variable.

$$logit\ \{\pi_{ij}\} = \gamma_0 + \gamma_1 sex + \gamma_2 edm + \gamma_3\left(opd_{ij} - \overline{opd}_{.j}\right) + \gamma_4(\overline{opd}_{.j} - \overline{opd}_{..}) \tag{2}$$

The negative binomial part of the model is conditioning the counts of negative responses to the students’ gender equality endorsement scale, the reverse sum scores denoting the rate of sexist responses.

$$log\ u_{ij} = \beta_0 + \beta_1 sex + \beta_2 edm + \beta_3\left(opd_{ij} - \overline{opd}_{.j}\right) + \beta_4\left(\overline{opd}_{.j} - \overline{opd}_{..}\right) \tag{3}$$

5. Results

Table 2 presents the unstandardized estimates of the fitted model. We describe the results in two parts. The first part is the ceiling part of the model. This part is a logistic regression, predicting students' endorsement of gender equality. The second part are estimates of a negative binomial model predicting the rate of responses of the reverse sum score. We interpret these estimates as endorsement of sexism because higher scores represent less agreement to assertions of endorsing gender equality and higher agreement to items expressing sexism.

Table 2. Zero-inflated negative binomial unstandardized estimates on zero scores of gender equality endorsement

	Sexism				Gender Equality			
	Remainder Negative binomial portion				Ceiling Logistic portion			
	E	SE	p <	EXP(E)	E	SE	p <	OR
Sex (girl)	-0.47	0.05	***	0.63	0.97	0.12	***	2.60
Mother education (tertiary)	-0.20	0.06	***	0.82	0.40	0.15	**	1.49
Open Classroom discussion (w)	-0.10	0.02	***	0.90	0.40	0.06	***	1.48
Open Classroom discussion (b)	-0.21	0.07	**	0.81	0.31	0.14	*	1.36
intercept	1.41	0.03	***	4.10	-1.77	0.00	***	
dispersion	0.38	0.04	***					

Note: E = unstandardized estimates, SE = standard errors, p<.05 *, p<.01 **, p<.001 ***; EXP(E) = exponentiated estimate or incidence rate ratio, OR = odds ratio.

5. 1 Ceiling part

All the chosen covariates are positively related to the endorsement of gender equality at the highest level. The odds of endorsement for girls are 2.60 times the odds of their male counterparts (E = 0.97, SE = 0.12, p < .001, OR = 2.60). Students whose mothers have tertiary education exhibit 1.49 times higher odds of endorsement than students whose mothers do not hold tertiary education credentials (E = 0.40, SE = 0.16, p < .01, OR = 1.49). Students in schools with a higher frequency of open classroom discussion (one standard deviation above the national mean) have 1.36 times higher odds of endorsement than those in schools with mean levels (E = 0.31, SE = 0.14, p < .05, OR = 1.36). Finally, students who report a higher frequency of open classroom discussions than their school peers have odds of endorsement 1.48 times that of their peers in the same school (E = 0.40, SE = 0.06, p < .001, OR = 1.48).

5. 2 Remainder part

While the ceiling part of the model is similar to logistic regression, the negative binomial part of the model provides estimates that, when exponentiated, can be interpreted as incident rate ratios ($e^{\beta} = IRR$). Because the reverse sum score of responses is not a count of events in the traditional sense (e.g., number of bullying events), we interpret the expected sexist score as if these were rates of sexist responses (i.e., number of sexist responses). Thus, an incidence rate of 18 implies that sexist responses were given to all items.

All the included covariates are negatively related to the rate of sexist responses. While holding all covariate values constant, among those students who do not have the highest level of gender equality of endorsement, the incidence rate of sexist responses for girls is .63 times the rate of sexist responses for boys (E = -0.47, SE = 0.05, p < .001, IRR = 0.63, 1/IRR = 1.59). The incidence rate of students with mothers with tertiary education is .82 times the incidence rate of students with mothers without tertiary education (E = -0.20, SE = 0.06, p < .001, IRR = 0.82, 1/IRR = 1.22). Students in schools with one standard deviation above Italy's national level of open classroom discussion present an incident rate .81 times the incident rate of students in schools with mean levels of open classroom discussion (E = -0.21, SE = 0.07, p < .01, IRR = 0.81, 1/IRR = 1.23). Finally, students who report higher levels of open classroom discussion present incident rates 0.90 times the incident rate of their classroom peers who report average levels of open classroom discussion (E = -0.10, SE = 0.02, p < .001, IRR = 0.90, 1/IRR = 1.11).

6. Conclusion and Discussion

The application of zero-inflated models enables a more nuanced interpretation of the relationship between covariates and response variables that exhibit floor or ceiling effects. This modeling approach provides two sub-models: a count model that takes care of the rate of responses and can account for long tails in the response variable; and a second sub-model that conditions the excess of zeroes that goes above the count model, where persons with the highest level of the attribute of interest may lay. By reversing the sum score of responses, this study illustrates how a zero-inflated model is applicable to sum scores subject to ceiling effects. The presented approach has precedents in the literature. Rutkowski et al. (2016) employed a multilevel zero-inflated Poisson model to analyze bullying responses, represented by a sum score. The novelty of the presented approach lies in the use of a reverse sum score, where the top score serves as a zero indicator.

We note two main limitations of the present approach: it cannot handle missing values in the outcome variable (only 64 cases were in this category in the current application), nor in the covariates (78 cases were missing in the selected covariates), and it does not account for measurement error. Missing values in the response variable are troublesome for all generalized linear models. The presented approach does not account for the measurement error inherent in the reverse sum score. This limitation is shared with any study that uses the partial credit score, mean score, or sum score of the responses to Likert-type instruments, such as the scale of endorsement of gender equality. Any of these scores, once conditioned in a generalized linear model does not have a term to account for measurement error. If the missing mechanism can be considered missing at random (Rubin, 1987), multiple imputations of the missing values could be done using a response model, such as the partial credit model at the item level. The variability of the plausible item responses given the partial credit model should carry the measurement error uncertainty. As such, if the multiple data sets are re-analyzed as in the present approach, it would be accounting for measurement error indirectly. In future research we plan to evaluate this later extension with simulation studies. Finally, alternative approaches such as factor mixture models can account for both zero inflation and measurement error (e.g., Wall et al., 2015; Carrasco et al., 2023). These approaches, however, involve greater model complexity than the ZINB framework adopted in the present study.

Ceiling and floor effects in scores are quite common among egalitarian attitudes scales. The inflation of cases at the ceiling of the distribution can underestimate the evaluation of protective factors and program intervention effects when linear regression models are used. While Tobit regression models may obtain corrected regressor coefficients for censored variables, these models are not informative about the relationship between covariates and the number of cases at the ceiling. Moreover, when evaluating regressors representing protective factors, school attributes, or interventions, the effect of interest may lie in the logit part of the model instead of the remainder (see Boulton et al., 2018). Zero-inflated negative binomial models are a more flexible option than zero-inflated Poisson if there is overdispersion. Finally, these models can be applied to studies with complex sample designs and expanded to multilevel models with random intercepts (Zhu et al., 2017). The present approach is a potential solution for the limitations linear regression models present when used to model response variables with ceiling or floor effects. The chosen model is expected to correct the underestimation of coefficients while also providing insight into which cases are more likely to reach the ceiling or floor. This application is of particular interest for scenarios where ceiling effects permit substantive interpretations, regardless of censoring. For example, interventions seeking to promote positive intergroup attitudes, such as those focused on gender equality, where scale scores are censored once students reach the maximum level of endorsement, especially when generating newer items expressing an even higher endorsement, might be too difficult.

Monte Carlo studies are needed to determine when zero-inflated models are more suitable than more popular models, such as linear regressions, and similar models such as two-part or hurdle models (Neelon et al., 2016). For example, what are the conditions that may hide effects of interest in regression models, such as the proportion of censored cases or variable overdispersion? Similarly, what are the more suitable conditions for applying the present approach in terms of sample size, number of items, and skewness? Thus, further research is needed to develop clear guidelines for the application of these models.

References

Asparouhov, T. (2005). Sampling Weights in Latent Variable Modeling. Structural Equation Modeling: A Multidisciplinary Journal, 12(3), 411–434. https://doi.org/10.1207/s15328007sem1203_4

Boulton, A. J., & Williford, A. (2018). Analyzing skewed continuous outcomes with many zeros: A tutorial for social work and youth prevention science researchers. Journal of the Society for Social Work and Research, 9(4), 721–740. https://doi.org/10.1086/701235

Carrasco, D., & Torres Irribarra, D. (2018). The Role of Classroom Discussion. In A. Sandoval-Hernández, M. M. Isac, & D. Miranda (Eds.), Teaching Tolerance in a Globalized World (Vol. 4, pp. 87–101). Springer International Publishing. https://doi.org/10.1007/978-3-319-78692-6_6

Carrasco, D., Torres Irribarra, D., & González, J. (2023). Continuation Ratio Model for Polytomous Responses with Censored Like Latent Classes. In M. Wiberg, D. Molenaar, J. González, J.-S. Kim, & H. Hwang (Eds.), Quantitative Psychology (pp. 243–256). Springer. https://doi.org/10.1007/978-3-031-27781-8_22

Green, J. A. (2021). Too many zeros and/or highly skewed? A tutorial on modelling health behaviour as count data with Poisson and negative binomial regression. Health Psychology and Behavioral Medicine, 9(1), 436–455. https://doi.org/10.1080/21642850.2021.1920416

Kang, H. A., Su, Y. H., & Chang, H. H. (2018). A note on monotonicity of item response functions for ordered polytomous item response theory models. British Journal of Mathematical and Statistical Psychology, 71(3), 523–535. https://doi.org/10.1111/bmsp.12131

Loeys, T., Moerkerke, B., de Smet, O., & Buysse, A. (2012). The analysis of zero-inflated count data: Beyond zero-inflated Poisson regression. British Journal of Mathematical and Statistical Psychology, 65(1), 163–180. https://doi.org/10.1111/j.2044-8317.2011.02031.x

López-Hornickel, N., Carrasco, D., Lay, S., & Treviño, E. (2024). It is not just your opinion. Gender equity endorsement of latin american students and their peers at school. Large-Scale Assessments in Education, 12(45). https://doi.org/10.1186/s40536-024-00235-6

McBee, M. (2010). Modeling Outcomes With Floor or Ceiling Effects: An Introduction to the Tobit Model. Gifted Child Quarterly, 54(4), 314–320. https://doi.org/10.1177/0016986210379095

Muthén, L. K., & Muthén, B. O. (2017). Mplus User's Guide (8th ed.). Los Angeles, CA: Muthén & Muthén.

Muthén, B. O., Muthén, L. K., & Asparouhov, T. (2016). Regression and Mediation Analysis using Mplus. Muthén & Muthén. https://statmodel.com/Mplus_Book.shtml

Neelon, B., O'Malley, A. J., & Smith, V. A. (2016). Modeling zero-modified count and semicontinuous data in health services research Part 1: background and overview. Statistics in Medicine, 35(27), 5070–5093. https://doi.org/10.1002/sim.7050

Rubin, D. B. (1987). Multiple Imputation for Nonresponse in Surveys. John Wiley & Sons.

Rutkowski, L., & Rutkowski, D. (2016). The Relation Between Students' Perceptions of Instructional Quality and Bullying Victimization. In Teacher Quality, Instructional Quality and Student Outcomes (Vol. 2, pp. 115–133). https://doi.org/10.1007/978-3-319-41252-8_6

Sandoval-Hernández, A., & Carrasco, D. (2020). A Measurement Strategy for SDG Thematic Indicators 4.7.4 and 4.7.5 Using International Large Scale Assessments in Education. http://tcg.uis.unesco.org/wp-content/uploads/sites/4/2020/06/Measurement-Strategy-for-474-and-475-using-ILSA_20200625.pdf

Schulz, W., Carstens, R., Losito, B., & Fraillon, J. (2018). ICCS 2016 Technical Report (W. Schulz, R. Carstens, B. Losito, & J. Fraillon (eds.)). International Association for the Evaluation of Educational Achievement (IEA).

Stapleton, L. M. (2008). Analysis of data from complex samples. In E. D. De Leeuw, J. J. Hox, & D. a. Dillman (Eds.), International Handbook of Survey Methodology (pp. 342–369). https://doi.org/10.4324/9780203843123_18

Wall, M. M., Park, J. Y., & Moustaki, I. (2015). IRT Modeling in the Presence of Zero-Inflation With Application to Psychiatric Disorder Severity. Applied Psychological Measurement, 39(8), 583–597. https://doi.org/10.1177/0146621615588184

Wilson, T., Loughran, T., & Brame, R. (2020). Substantial Bias in the Tobit Estimator: Making a Case for Alternatives. Justice Quarterly, 37(2), 231–257. https://doi.org/10.1080/07418825.2018.1517220

Zhu, L., & Gonzalez, J. (2017). Modeling floor effects in standardized vocabulary test scores in a sample of low SES Hispanic preschool children under the multilevel structural equation modeling framework. Frontiers in Psychology, 8(DEC), 1–14. https://doi.org/10.3389/fpsyg.2017.02146