

APPLICATION NOTE



## Estimating intraclass correlation for ordinal data

Benjamin W. Langworthy<sup>a,b</sup>, Zhaoxun Hou<sup>a</sup>, Gary C. Curhan<sup>b,c,d</sup>, Sharon G. Curhan<sup>c</sup> and Molin Wang<sup>a,b,c</sup>

<sup>a</sup>Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA, USA; <sup>b</sup>Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, MA, USA; <sup>c</sup>Department of Medicine, Channing Division of Network Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA, USA; <sup>d</sup>Renal Division, Brigham and Women's Hospital, Boston, MA, USA

### ABSTRACT

In this paper, we consider the estimation of intraclass correlation for ordinal data. We focus on pure-tone audiometry hearing threshold data, where thresholds are measured in 5 decibel increments. We estimate the intraclass correlation for tests from iPhone-based hearing assessment applications as a measure of test/retest reliability. We present a method to estimate the intraclass correlation using mixed effects cumulative logistic and probit models, which assume the outcome data are ordinal. This contrasts with using a mixed effects linear model which assumes that the outcome data are continuous. In simulation studies, we show that using a mixed effects linear model to estimate the intraclass correlation for ordinal data results in a negative finite sample bias, while using mixed effects cumulative logistic or probit models reduces this bias. The estimated intraclass correlation for the iPhone-based hearing assessment application is higher when using the mixed effects cumulative logistic and probit models compared to using a mixed effects linear model. When data are ordinal, using mixed effects cumulative logistic or probit models reduces the bias of intraclass correlation estimates relative to using a mixed effects linear model.

### ARTICLE HISTORY



Received 20 December 2022  
Accepted 1 November 2023

### KEYWORDS

Test/retest reliability;  
reliability and validity;  
pure-tone audiometry;  
intraclass correlation;  
ordinal data

## Introduction

Intraclass correlation (ICC) is frequently used to measure test/retest reliability and can be used as one tool to measure the quality of newly developed testing procedures [3]. We consider audiometric hearing threshold data from an iPhone-based hearing assessment application. Pure-tone audiometry assesses the quietest tone, 'hearing threshold,' that can be detected at specific frequencies in each ear. If tests are conducted on the same subject in a short enough time frame we expect the true underlying hearing threshold to remain constant for each test. Therefore, a reliable test should have a high ICC, indicating a strong correlation among the hearing thresholds of the same ear.

**CONTACT** Benjamin W. Langworthy  langw019@umn.edu  Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA, USA; Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, MA, USA

ICC can be estimated using a linear mixed effects model which assumes the random effects and random error terms are normally distributed. However, pure tone audiometry tests typically measure hearing thresholds in 5 decibel (dB) increments. Therefore, hearing threshold data are ordinal rather than continuous. In simulations, we show that for finite samples, treating the data as continuous, rather than ordinal, biases the estimates of the ICC toward zero. A framework for estimating the ICC for generalized linear mixed effects models has been proposed in Refs [5,6,11,13,14]. We extend this framework to cumulative link models for ordinal data. We show that using this ordinal model framework can obtain more accurate ICC estimates than naively treating ordinal data as continuous.

## Methods

Let  $Y_{ij}^*$  denote the true reading for the  $j$ th measurement for the  $i$ th cluster on a continuous scale. Assume that  $Y_{ij}^*$  follows a linear mixed effects model,

$$Y_{ij}^* = X_{ij}^T \beta^* + b_i^* + \epsilon_{ij}^*, \quad (1)$$

where  $X_{ij}$  is a vector of covariates,  $\beta^*$  is a vector of fixed effect parameters,  $b_i^* \sim N(0, \sigma_{b^*}^2)$  is an ear-specific random effect, and  $\epsilon_{ij}^* \sim N(0, \sigma_{\epsilon^*}^2)$  is a normally distributed random error term. We focus on the adjusted ICC [4,14], which is

$$ICC_{Y_{ij}^*}^{adj} = \frac{\sigma_{b^*}^2}{\sigma_{b^*}^2 + \sigma_{\epsilon^*}^2}.$$

We focus on the adjusted ICC because this allows us to estimate the ICC after controlling for covariates,  $X_{ij}$ , which may affect the outcome but are not of primary interest for the study. As an example, we adjust for mean ambient noise in our illustrative example because it may change the measured hearing threshold for reasons other than test/retest reliability. For ordinal data, instead of observing  $Y_{ij}^*$  we observe  $Y_{ij}$ , which has  $K$  categories. In order to categorize  $Y_{ij}$ , we assume there are  $K + 1$  different cutpoints,  $\xi_k$ , where  $Y_{ij} = k$  if  $\xi_{k-1} < Y_{ij}^* < \xi_k$ . In this case  $\xi_0 = -\infty$  and  $\xi_K = \infty$ , and all the other cutpoints are finite values put into ascending order. This can be written as a mixed effects cumulative probit model,

$$\begin{aligned} P(Y_{ij} \leq k) &= P(Y_{ij}^* \leq \xi_k) = \Phi \left( \frac{\xi_k - X_{ij}^T \beta^* - b_i^*}{\sigma_{\epsilon^*}} \right) \\ &= \Phi \left( \frac{\xi_k}{\sigma_{\epsilon^*}} - \frac{X_{ij}^T \beta^*}{\sigma_{\epsilon^*}} - \frac{b_i^*}{\sigma_{\epsilon^*}} \right), \end{aligned}$$

where the fixed effects coefficients are  $\beta = \beta^* / \sigma_{\epsilon^*}$ , and the random effects are  $b_i = b_i^* / \sigma_{\epsilon^*}$ , with  $b_i \sim N(0, \sigma_b^2)$ , and  $\sigma_b^2 = \sigma_{b^*}^2 / \sigma_{\epsilon^*}^2$  [1,10]. We can recover the adjusted ICC for the latent variable,  $Y_{ij}^*$ , as

$$ICC_{Y_{ij}^*}^{adj} = \frac{\sigma_b^2}{\sigma_b^2 + 1}.$$

If we assume that  $\epsilon_{ij}^*$  has a logistic distribution, the same method would imply a mixed effects cumulative logistic model. In this case, when calculating the adjusted ICC, we

replace the one in the denominator with  $\pi^2/3$ , the variance of a standard logistic distribution,

$$ICC_{Y_{ij}^* \log}^{adj} = \frac{\sigma_b^2}{\sigma_b^2 + \pi^2/3}.$$

Above we have a single level of clustering. It is also possible to have two levels of clustering, with one nested within the other. For instance, in our illustrative example with hearing data, we have one level of clustering due to individuals being tested multiple times, and a nested cluster due to each individual having two ears. Assume that  $Y_{ikj}^*$  is the underlying true reading on a continuous scale of the  $j$ th measurement for the  $i$ th first-level cluster and the  $k$ th second-level cluster. Assume  $Y_{ikj}^*$  follows the linear mixed effects model,

$$Y_{ikj}^* = X_{ikj}^T \beta^* + b_i^* + c_{ik}^* + \epsilon_{ikj}^*, \quad (2)$$

where  $b_i^* \sim N(0, \sigma_{b^*}^2)$  is a first-level cluster-specific random effect,  $c_{ik}^* \sim N(0, \sigma_{c^*}^2)$  is a second-level cluster-specific random effect, and  $\epsilon_{ikj}^* \sim N(0, \sigma_{\epsilon^*}^2)$  is a random error term. The ICC for two measurements from the same first and second-level clusters is defined as  $ICC_{Y_{ikj}^*}^{adj} = \text{Cor}(Y_{ikj}^*, Y_{ikj'}^*) = \frac{\text{Cov}(Y_{ikj}^*, Y_{ikj'}^*)}{\sqrt{\text{Var}(Y_{ikj}^*)} \sqrt{\text{Var}(Y_{ikj'}^*)}}$ , for  $j \neq j'$ . Based on Equation (2), this can be rewritten as

$$ICC_{Y_{ikj}^*}^{adj} = \frac{\sigma_{b^*}^2 + \sigma_{c^*}^2}{\sigma_{b^*}^2 + \sigma_{c^*}^2 + \sigma_{\epsilon^*}^2}.$$

Using the same methods as the single random effect model, the adjusted ICC based on the mixed effects cumulative probit and logistic models can be written as

$$ICC_{Y_{ikj}^* \text{prob/log}}^{adj} = \frac{\sigma_b^2 + \sigma_c^2}{\sigma_b^2 + \sigma_c^2 + m},$$

where  $m = 1$  for the mixed effects cumulative probit model,  $m = \pi^2/3$  for the mixed effects cumulative logistic model,  $\sigma_b = \sigma_{b^*}/\sigma_{\epsilon^*}$ , and  $\sigma_c = \sigma_{c^*}/\sigma_{\epsilon^*}$ .

The adjusted ICC for ordinal data can be estimated by plugging in the variance estimates for the random effects into the formulas for  $ICC_{Y_{ij}^* \text{prob/log}}^{adj}$  or  $ICC_{Y_{ijk}^* \text{prob/log}}^{adj}$ . We will denote the estimates as  $\widehat{ICC}_{Y_{ij}^* \text{prob/log}}^{adj}$  and  $\widehat{ICC}_{Y_{ijk}^* \text{prob/log}}^{adj}$ . It is also of interest to obtain confidence intervals to quantify the uncertainty of these estimates. For the single random effect setup we note that  $ICC_{Y_{ij}^* \text{prob}}^{adj}$  and  $ICC_{Y_{ij}^* \log}^{adj}$  are both transformations of  $\sigma_b^2$ . Therefore we can calculate the confidence interval for  $ICC_{Y_{ij}^* \text{prob}}^{adj}$  and  $ICC_{Y_{ij}^* \log}^{adj}$  by transforming the confidence interval for  $\sigma_b^2$ , which can be obtained using the likelihood root statistic from the profile likelihood [8,9].

In the multiple random effect setting we can obtain a confidence interval for  $\widehat{ICC}_{Y_{ijk}^* \text{prob}}^{adj}$  and  $\widehat{ICC}_{Y_{ijk}^* \log}^{adj}$  by applying the delta method based on the estimated variance/covariance

matrix of  $\sigma_b^2$  and  $\sigma_c^2$ . If

$$\begin{bmatrix} \hat{\sigma}_b \\ \hat{\sigma}_c \end{bmatrix} \rightarrow_d N(0, \theta),$$

then by the delta method

$$\widehat{ICC}_{Y_{ijk}^*}^{adj} \rightarrow_d N[0, g'(\sigma_b, \sigma_c, m)^T \theta g'(\sigma_b, \sigma_c, m)],$$

where

$$g'(\sigma_b, \sigma_c, m) = \begin{bmatrix} \frac{2 \cdot \sigma_b \cdot m}{(\sigma_b^2 + \sigma_c^2 + m)^2} \\ \frac{2 \cdot \sigma_c \cdot m}{(\sigma_b^2 + \sigma_c^2 + m)^2} \end{bmatrix}.$$

We include an example of how to estimate the adjusted ICC and confidence intervals for all methods at <https://github.com/blangworthy/ICCordinal>.

## Application

### Simulation study

We use simulation studies to compare the finite sample performance of the adjusted ICC estimates using mixed effects cumulative probit or logistic models to those using mixed effects linear regression, when the simulated data are ordinal. First we consider a single level of clustering, where the continuous outcome  $Y_{ij}^*$  is simulated according to the mixed effects linear model from Equation (1), where  $X_{ij}$  is a scalar covariate,  $\beta^*$  is the fixed effect parameter,  $b_i^*$  is an ear-specific random effect, and  $\epsilon_{ij}^*$  is a random error term. We simulate 35 clusters with 5 measurements per cluster, such that  $i = 1, \dots, 35$  and  $j = 1, \dots, 5$ . We simulate  $X \sim N(0, 1)$ ,  $\beta^* = 1$ , and  $b_i^* \sim N(0, 4)$ . For the random error term we simulate both  $\epsilon^* \sim N(0, 1)$  and  $\epsilon^* \sim \text{Logistic}(0, 2\sqrt{3}/\pi)$ , with the former corresponding to a properly specified cumulative probit model and the latter corresponding to a properly specified cumulative logistic model after using cutpoints to define an ordinal variable based on  $Y_{ij}^*$ . In both cases, the true adjusted ICC is 0.8. We consider three different ways to define the ordinal variable,  $Y_{ij}$ . The first, which we will refer to as setup one, is to define the cutpoints,  $\xi_k$ , as all even integers including 0. This results in 7–9 categories for most simulations. For another setup, which we will refer to as setup two, we define the cutpoints,  $\xi_k$ , as  $-\infty, -2, 0, 2$ , and  $\infty$ , which results in four categories. For our final setup, which we will refer to as setup three, we define the cutpoints as  $-\infty, -2, 1, 2$ , and  $\infty$ . Setup one has a similar number of categories to our real data example, while setups two and three are used to show how results change when there are a smaller number of categories. In order to estimate the ICC we fit mixed effects cumulative probit and logistic models and use the methods described in the Methods section. We also use the naive method, in which we fit a mixed effects linear model with the ordinal variable,  $Y_{ij}$ , as the outcome. We note that the ordinal values,  $Y_{ij}$ , are all evenly spaced, even in the case of setup three when the cutpoints are not evenly spaced.

In addition to the single level of clustering, we also consider simulations using two levels of clustering where we simulate  $Y_{ikj}^*$  according to the linear mixed effects model from Equation (2), where  $b_i^*$  is a subject-specific random effect,  $c_{ik}^*$  is an ear specific random

**Table 1.** Bias (SD) and 95% confidence interval coverage rate of adjusted ICC estimates for single and multi-level clustering for the three different setups using three different methods for estimation.

	Single-level clustering <sup>a</sup>				Multi-level Clustering <sup>a</sup>			
	Normal error <sup>b</sup>		Logistic error <sup>b</sup>		Normal error		Logistic error	
	Bias (SD) <sup>c</sup>	CR <sup>d</sup>	Bias (SD)	CR	Bias (SD)	CR	Bias (SD)	CR
Setup one <sup>e</sup>								
Adjusted ICC Probit	−0.01 (0.05)	0.95	−0.01 (0.06)	0.92	−0.01 (0.04)	0.93	−0.01 (0.04)	0.92
Adjusted ICC Logit	−0.01 (0.06)	0.95	−0.01 (0.06)	0.93	−0.01 (0.04)	0.94	−0.01 (0.04)	0.92
Adjusted ICC Naïve	−0.06 (0.06)	0.89	−0.06 (0.06)	0.87	−0.06 (0.04)	0.80	−0.06 (0.04)	0.81
Setup two <sup>f</sup>								
Adjusted ICC Probit	−0.01 (0.06)	0.95	−0.01 (0.06)	0.95	−0.01 (0.04)	0.94	−0.01 (0.04)	0.94
Adjusted ICC Logit	−0.02 (0.06)	0.96	−0.01 (0.06)	0.94	−0.02 (0.04)	0.95	−0.01 (0.04)	0.97
Adjusted ICC Naïve	−0.11 (0.06)	0.68	−0.10 (0.06)	0.70	−0.10 (0.04)	0.40	−0.10 (0.04)	0.43
Setup three <sup>g</sup>								
Adjusted ICC Probit	−0.01 (0.06)	0.95	−0.01 (0.06)	0.94	−0.01 (0.04)	0.94	−0.01 (0.04)	0.95
Adjusted ICC Logit	−0.02 (0.06)	0.96	−0.02 (0.06)	0.94	−0.02 (0.04)	0.95	−0.01 (0.04)	0.95
Adjusted ICC Naïve	−0.14 (0.06)	0.48	−0.14 (0.07)	0.50	−0.13 (0.05)	0.19	−0.13 (0.05)	0.20

Note: True ICC = 0.8.

<sup>a</sup>Single-level clustering setup includes 35 subjects and 5 measures per subject. Multi-level clustering setup includes 35 subjects, 2 ears per subject, and 5 measures per ear.

<sup>b</sup>When error has normal distribution probit model is properly specified. When error has logistic distribution logistic model is properly specified.

<sup>c</sup>Bias is the estimated ICC minus the true ICC; SD is the empirical standard deviation of ICC estimates over all simulation replicates.

<sup>d</sup>CR is coverage rate and is the proportion of 95% confidence intervals that contain the true ICC; confidence intervals for adjusted ICC using probit and logistic models in single-level clustering use profile confidence interval.

<sup>e</sup>Setup one sets cutpoints at all even integers including 0, resulting in 7–9 categories in most simulations

<sup>f</sup>Setup two sets cutpoints at −2, 0 and 2, resulting in 4 categories

<sup>g</sup>Setup three sets cutpoints at −2, 1 and 2, resulting in 4 categories

effect, and  $\epsilon_{ikj}^*$  is a random error term. We simulate such that  $i = 1, \dots, 35$ ,  $k = 1, 2$  and  $j = 1, \dots, 5$ . This is equivalent to a study where there are 35 first-level clusters with 2 second-level clusters per first-level cluster each, and 5 measurements per first- and second-level cluster. We define  $X$ ,  $\beta^*$  and  $\epsilon^*$  in the same way as the single-level clustering example. We simulate  $b_i^* \sim N(0, 2)$ , and  $c_{ik}^* \sim N(0, 2)$ , so that the true ICC is again 0.8. Again, we define the ordinal variable  $Y_{ikj}$  using the same three setups, with the same cutpoints as used to define  $Y_{ij}$  in the single level of clustering simulations. As with the single clustering simulations we estimated the ICC using methods described in the Methods section. We compare using mixed effects cumulative logistic and probit models to a mixed effects linear model, with  $Y_{ikj}$  as the outcome.

The empirical bias, standard deviation, and 95% confidence interval coverage rate based on 1000 simulations for both the single and multi-level clustering setups are reported in Table 1. The adjusted ICC estimates using the mixed effect cumulative logistic and probit models both have bias close to zero while the naive estimator underestimates the adjusted ICC. This negative bias for the naive estimator gets larger for setups two and three where there are fewer categories. This suggests that the naive estimator performs even more poorly in cases where there are fewer categories. In particular, for setup three, where the cutpoints defining the categories are not evenly spaced the naive estimator does even more poorly. There is also an issue of the confidence interval not being able to be estimated because the cumulative logistic and probit models are unstable when the estimated ICC is one, or very close to one. This is relatively rare (2.6% of simulated data sets or less), except

for the cumulative logistic model for single-level clustering for setups two and three. For setup two and three, in the single level clustering scenario, the confidence interval for the cumulative logistic model could not be estimated for between 12 and 13% of the simulated data sets. When the ICC is one the variance of  $\epsilon_{ij}^*$  or  $\epsilon_{ijk}^*$  is zero, and the cumulative logistic or probit models will not be well defined due to division by zero. This issue can cause issues with the point estimate, however, our simulated datasets only caused issue with defining the confidence intervals. This can be due to the limits of the profile confidence interval not being defined. Alternatively the variance/covariance matrix for the random effects not being positive semi-definite, which will lead to the Wald confidence interval to not be well defined. We exclude the simulated data sets where the confidence intervals could not be estimated when calculating the coverage rate. The confidence intervals have close to the desired coverage rate when using the cumulative logistic or probit models.

### **Illustrative example**

To illustrate our method, we analyze data from 31 women from the Nurses' Health Study II (NHS II) [2], who self-administered home hearing tests using the Decibel Therapeutics iPhone hearing assessment application. The home hearing test assessed pure-tone air conduction hearing thresholds (in 5 dB steps) across the conventional range of frequencies between 500 and 8000 hertz (Hz) in each ear. In the study 21 (68%) of the participants had measurements from four repeated home hearing tests; 27 (87%) had measurements from at least two repeated tests. The average number of home hearing assessments was 3.38. The results for single-ear and both-ear adjusted ICCs for hearing threshold measurements obtained at each of the frequencies are reported in Table 2. For both the single-ear and both-ear adjusted ICCs we adjust for mean ambient noise measured in decibels. In general, the adjusted ICC estimated using the mixed effects cumulative logistic model is the highest, and the adjusted ICC estimated using the mixed effects linear model is the lowest. This is consistent with simulation results, in which the naive estimator using the mixed effects linear model has a negative finite sample bias. The adjusted ICC is above 0.7 in all cases, across all frequencies. For the both ear adjusted models using the cumulative logistic and probit models the lower bound of the 95% confidence interval is also above 0.7 for all frequencies except for 2000 Hz under the cumulative probit model. Further the lower bound of the 95% confidence interval is above 0.5 for all methods and frequencies except for the naive method and the cumulative probit model for 2000 Hz for the left ear. This indicates strong test/retest reliability. One thing we note is that for the adjusted ICC using both ears at 4000 Hz cannot be estimated using the cumulative logistic model. This is due to the issue with the estimated adjusted ICC being one causing issues due to division by zero, as discussed in Section 3 of the main text. There are a number of frequencies where the confidence intervals cannot be estimated due to a similar issue.

### **Conclusion and discussion**

We show that when data are ordinal, using a linear mixed effects model which treats data as continuous can underestimate the adjusted ICC. This is particularly true when there are a small number of categories in ordinal data, or the spacing between categories is uneven.

**Table 2.** Single-ear adjusted ICCs and 95% Wald and profile confidence intervals for all 500, 1000, 2000, 3000, 4000, 6000, and 8000 Hz.

	Cumulative probit		Cumulative logistic		Naive	
	Adj ICC	95% CI <sup>a</sup>	Adj ICC	95% CI	Adj ICC	95% CI
	Frequency/ear					
500 Hz/Right	0.80	(0.63,0.90)	0.81	NA	0.80	(0.69,0.92)
500 Hz/Left	0.92	NA	0.93	(0.84,0.97)	0.89	(0.83,0.96)
1000 Hz/Right	0.83	(0.69,0.91)	0.86	(0.73,0.93)	0.77	(0.64,0.90)
1000 Hz/Left	0.77	(0.57,0.88)	0.79	NA	0.73	(0.57,0.88)
2000 Hz/Right	0.80	(0.65,0.90)	0.90	NA	0.78	(0.66,0.91)
2000 Hz/Left	0.60	(0.34,0.79)	0.75	(0.52,0.88)	0.57	(0.36,0.79)
3000 Hz/Right	0.78	(0.60,0.89)	0.83	(0.67,0.92)	0.75	(0.61,0.89)
3000 Hz/Left	0.99	NA	0.99	NA	0.97	(0.95,0.99)
4000 Hz/Right	0.75	(0.56,0.88)	0.85	(0.70,0.94)	0.84	(0.74,0.94)
4000 Hz/Left	0.99	NA	1.00	NA	0.98	(0.96,0.99)
6000 Hz/Right	0.85	(0.72,0.93)	0.89	(0.78,0.95)	0.80	(0.69,0.92)
6000 Hz/Left	0.92	NA	0.93	(0.83,0.97)	0.93	(0.88,0.98)
8000 Hz/Right	0.92	(0.84,0.96)	0.94	(0.87,0.98)	0.89	(0.81,0.96)
8000 Hz/Left	0.86	(0.75,0.93)	0.91	(0.82,0.96)	0.86	(0.78,0.95)
500 Hz/Both	0.85	(0.78,0.93)	0.87	(0.79,0.94)	0.85	(0.77,0.92)
1000 Hz/Both	0.81	(0.71,0.91)	0.84	(0.75,0.93)	0.76	(0.65,0.87)
2000 Hz/Both	0.73	(0.60,0.85)	0.84	(0.76,0.92)	0.70	(0.57,0.83)
3000 Hz/Both	0.87	(0.80,0.93)	0.92	(0.88,0.97)	0.85	(0.78,0.92)
4000 Hz/Both	0.92	(0.87,0.96)	NA	NA	0.91	(0.87,0.96)
6000 Hz/Both	0.88	(0.81,0.95)	0.90	(0.84,0.96)	0.85	(0.78,0.93)
8000 Hz/Both	0.89	(0.83,0.95)	0.93	(0.88,0.97)	0.88	(0.82,0.94)

Note: Mixed effects model included mean ambient noise as covariate.

NA's produced when issue created due to division by zero.

<sup>a</sup>Confidence intervals for single-ear adjusted cumulative logistic and probit models are estimated using transformed profile confidence intervals, all other confidence intervals estimated using the delta method.

A more precise estimate of the adjusted ICC can be obtained using a mixed effects cumulative logistic or probit model. One potential limitation to these methods is if neither mixed effects cumulative logistic or probit models fit the data well, alternative methods may be necessary. Previous work estimated the ICC for binary data with a mixed effects logistic regression model using a similar latent variable approach, with data simulated from a beta-binomial distribution and showed that performance varied depending on the parameters of the simulated distribution [7]. Other work, which also simulated data using the multivariate normal latent variable approach, showed strong performance using a generalized estimating equation approach to calculate the ICC for ordinal data [12], although the method presented in that paper is only applicable for balanced data with an equal number of replicates for each subject or rater.

We apply our methods to hearing threshold data obtained from participants in the NHS II. In this example, the estimated adjusted ICC using mixed effects cumulative logistic or probit models is higher than the estimated adjusted ICC using a linear mixed effects model. Example code for estimating the adjusted ICC can be found at <https://github.com/blangworthy/ICCordinal>.

When data are ordinal rather than continuous it is best to estimate the ICC using ordinal mixed effects models, rather than mixed effects linear models. This will lead to more accurate results, as the ICC estimate from the mixed effects linear model tends to be attenuated for ordinal data. Our simulations and results indicate this is particularly true when



there are a small number of categories, but are still the case for ordinal data with as many as 7–9 categories.

## Ethics approval and consent to participate

The study protocol was approved by the institutional review board of the Brigham and Women's Hospital, Boston, MA. Return of self-administered questionnaire was considered as implied consent by the institutional review boards. All participants in the CHEARS Audiometry Assessment Arm provided written informed consent.

## Disclosure statement

SGC serves as a consultant to Decibel Therapeutics. GCC is an employee of OM1, has equity in Allena Pharmaceuticals, and receives royalties from UpToDate for being an author and Section Editor. All other authors have no competing interests to report.

## Funding

This work was supported by the National Institute Health [grant numbers R01 DC017717 and U01 DC010811].

## Data availability statement

The Nurses' Health Study (NHS) II supports transparency and has data sharing mechanisms clearly in place, which have been approved by their IRB. Further details can be found at <http://www.nurseshealthstudy.org/researchers> and <http://www.nurseshealthstudy.org/contact>. Further questions about the data can be addressed to [nhsaccess@channing.harvard.edu](mailto:nhsaccess@channing.harvard.edu). The CHEARS study can be contacted at [CHEARS@channing.harvard.edu](mailto:CHEARS@channing.harvard.edu).

Relevant code can be found at <https://github.com/blangworthy/ICCordinal>

## References

- [1] A. Agresti, *Categorical Data Analysis*, Hoboken, NJ: John Wiley & Sons, 2003.
- [2] Y. Bao, M.L. Bertoia, E.B. Lenart, M.J. Stampfer, W.C. Willett, F.E. Speizer, and J.E. Chavarro, *Origin, methods, and evolution of the three nurses' health studies*, Am. J. Public. Health. 106 (2016), pp. 1573–1581.
- [3] J. Bartko, *The intraclass correlation coefficient as a measure of reliability*, Psychol. Rep. 19 (1966), pp. 3–11.
- [4] W. Browne, S. Subramanian, K. Jones, and H. Goldstein, *Variance partitioning in multilevel logistic models that exhibit overdispersion*, J. R. Stat. Soc. Ser. A: Stat. Soc. 168 (2005), pp. 599–613.
- [5] J. Carrasco and L. Jover, *Concordance correlation coefficient applied to discrete data*, Stat. Med. 24 (2005), pp. 4021–4034.
- [6] J. Carrasco, *A generalized concordance correlation coefficient based on the variance components generalized linear mixed models for overdispersed count data*, Biometrics 66 (2010), pp. 897–904.
- [7] J. Carrasco, Y. Pan, and R. Abellana, *Estimating marginal proportions and intraclass correlations with clustered binary data*, Biometrical J. 61 (2019), pp. 574–599.
- [8] R.H.B. Christensen, *Analysis of ordinal data with cumulative link models—estimation with the R-package ordinal*, R-package Version Vol. 28, 2015, pp. 406.
- [9] R.H.B. Christensen, *Cumulative link models for ordinal regression with the R package ordinal*, J. Stat. Softw. (2018). Available at online: <http://www.nurseshealthstudy.org/researchers> and [https://cran.uni-muenster.de/web/packages/ordinal/vignettes/clm\\_article.pdf](https://cran.uni-muenster.de/web/packages/ordinal/vignettes/clm_article.pdf)



- [10] D. Hedeker and R.D. Gibbons, *A random-effects ordinal regression model for multilevel analysis*, Biometrics 5 (1994), pp. 933–944.
- [11] S. Nakagawa and H. Schielzeth, *Repeatability for gaussian and non-gaussian data: A practical guide for biologists*, Biol. Rev. 85 (2010), pp. 935–956.
- [12] L. Lin, A. Hedayat, and W. Wu, *A unified approach for assessing agreement for continuous and categorical data*, J. Biopharm. Stat. 17 (2007), pp. 629–652.
- [13] S. Nakagawa and H. Schielzeth, *A general and simple method for obtaining  $r^2$  from generalized linear mixed-effects models*, Methods Ecol. Evol. 4 (2013), pp. 133–142.
- [14] S. Nakagawa, P.C. Johnson, and H. Schielzeth, *The coefficient of determination  $r^2$  and intra-class correlation coefficient from generalized linear mixed-effects models revisited and expanded*, J. R. Soc. Interface 14 (2017), pp. 20170213.