Comparing Pseudo-R-squared Metrics for Multilevel Logistic Regression Models

Zixie Zheng

A thesis

submitted in partial fulfillment of the

requirements for the degree of

Master of Education

University of Washington

2025

Committee:

Elizabeth Sanders

Oscar Olvera Astivia

Program Authorized to Offer Degree:

Education

University of Washington

Abstract

Comparing Pseudo-R-squared Metrics for Multilevel Logistic Regression Models

Zixie Zheng

Chair of the Supervisory Committee:

Elizabeth Sanders

Education

Reporting effect sizes is an important staple of modern research, as they facilitate the quantification of variable relations for meta-analyses irrespective of operational scales used, enable policymakers to understand practical returns on a program investment, help alleviate reliance on null hypothesis testing, and are an essential ingredient for planning research sample sizes. In single-level logistic regression, it is well-known that traditional variance-explained metrics like those in linear regression are not possible; instead, pseudo-R^2 measures that capture how much better a model with regressors fits compared to a null model are in wide use and are available in popular statistical computing packages. While Nakagawa et al. (2017) have proposed a McKelvey-Zavoina-type metric, there have been no systematic studies of effect sizes for multilevel (random effects) logistic regression models. Further, our systematic review of educational and psychological applied literature over the past five years indicates that most research employing multilevel logistic models does not report any effect sizes – likely due to the gap in the methodological literature. The present study therefore uses

simulation to compare the performance of pseudo-R^2 metrics for multilevel logistic regression models, including the McKelvey-Zavoina proposed by Nakagawa, as well as a new adjusted McKelvey-Zavoina metric, for 2-level random intercept models with varying ICCs, numbers of clusters, and cluster sizes. Those results indicate that the adjusted-McKelvey-Zavoina metric was best at reproducing the underlying data-generating R^2 value. Limitations and future research directions are discussed.

Keywords: Multilevel logistic regression, Pseudo-R^2, Effect size, Generalized linear model, Random effects models

Comparing Pseudo-R-squared Metrics for Multilevel Logistic Regression Models

In modern social science research, statistical significance testing has long been the primary method for evaluating research results. However, as research methodology has evolved, relying solely on null hypothesis testing *p*-values has been widely recognized as insufficient (Wilkinson, 1999). *P*-values merely indicate the compatibility of observed data with a specific null hypothesis, but fail to provide direct information about effect size, practical significance, or the importance of results. Consequently, reporting effect sizes has become a hallmark of high-quality research, not only providing information about the actual magnitude of effects but also facilitating the comparability of research results, the conduct of meta-analyses, and the assessment of practical significance (Cohen, 1988; Fritz et al., 2012; Thompson, 2007).

Despite the widely recognized importance of effect size reporting, implementing this practice in complex models can present challenges. In linear models, R^2 and its variants (such as adjusted R^2 and partial R^2) are widely accepted as standard effect size indicators. The advantages of R^2 lie in its intuitive nature, standardized characteristics (range from 0 to 1), and widely accepted interpretative framework (e.g., Cohen, 1988). Nevertheless, when research is focused on a categorical outcome, the traditional R^2 cannot be used as an effect size because the dependent variable in generalized linear models such as logistic regression is a nonlinear function of probability rather than a continuous variable (Menard, 2000). Specifically, in linear regression, the deviations (residuals) between model predictions and observed values are homogeneous and can be represented by error variance. In contrast, in logistic regression, the conditional variance of the dependent variable is a function of its predicted probabilities, which complicates the concept of "proportion of explained variance." Furthermore, parameter estimation in logistic regression is obtained through maximum likelihood methods rather than least squares methods, which further increases the difficulty of

directly applying traditional R^2 (Long, 1997). This fundamental difference leads to complexity in effect size calculation and interpretation, making it difficult for researchers to use simple, unified effect size indicators when dealing with binary outcome variables.

Pseudo-R^2 Indicators in Single-Level Logistic Regression

To address the challenge of representing effect sizes in binary outcome models, statisticians have developed various pseudo-R^2 metrics to mimic the traditional R^2 metrics used in linear regression (Menard, 2000; Long, 1997). Nevertheless, these metrics differ from the traditional R^2 in their theoretical foundations, calculation methods, and resulting interpretations (e.g., Hosmer et al., 2000; Long, 1997). Below I review the four most common pseudo-R^2 metrics.

McFadden's pseudo-R^2 (McFadden, 1974) metric is based on the comparison of two model likelihoods, as follows:

$$R^2_{McF} = 1 - \left(\frac{ln(L_{full})}{ln(L_{null})}\right) = 1 - \left(\frac{LL_{full}}{LL_{null}}\right), \quad (1)$$

where L_{full} is the likelihood of the "full" model (i.e., with predictors) and L_{null} is the likelihood of the "null" model (intercept-only without predictors, also known as a "baseline" model). Its theoretical foundation comes from information theory and can be interpreted as the proportion of information loss reduced by including predictor variables in the model. Although its theoretical upper limit approaches but does not equal 1, McFadden considered values greater than .20 to indicate a "good" fit and values greater than .40 to indicate "very good" fit. The advantage of this metric lies in its theoretical foundation and simple calculation, but its disadvantage is that its values are typically lower in magnitude, with values of .20 – .40 considered excellent fit, compared to the familiar scale of traditional R^2 (McFadden, 1979).

Like the R^2_{McF}, the **Cox & Snell pseudo-R^2** (Cox & Snell, 1989) is also based on a comparison of two likelihoods, but makes an adjustment for sample size, as follows.

$$R_{CS}^2 = 1 - \left(\frac{LL_{null}}{LL_{full}}\right)^{\frac{2}{n}} = 1 - exp^{\left((2/n)\times\left(LL_{null} - LL_{full}\right)\right)} = 1 - exp^{\left(\frac{2\left(LL_{null} - LL_{full}\right)}{n}\right)} \quad (2)$$

In (2), n is the total sample size, LL_{full} is the log-likelihood of the full model, and LL_{null} is the log-likelihood of the null (baseline) model. Its theoretical foundation derives from the likelihood function and can be interpreted as the proportion of maximum possible improvement achieved by introducing predictor variables. The main limitation of this metric is that its theoretical upper limit depends on the baseline model probability of the dependent variable, making cross-study comparisons difficult.

The **Nagelkerke pseudo-R^2** (Nagelkerke, 1991) is an adjusted version of R_{CS}^2 that is standardized by dividing R_{CS}^2 by the null (baseline) model's probability, allowing the metric's upper limit to reach 1, as follows.

$$R_N^2 = \frac{1 - exp^{\left(\frac{2\left(LL_{null} - LL_{full}\right)}{n}\right)}}{1 - exp^{\left(\frac{2\left(LL_{null}\right)}{n}\right)}} \quad (3)$$

The R_N^2 is arguably one of the most widely used pseudo-R^2 metrics in applied social science research and is routinely reported in popular software for generalized linear models.

The **McKelvey-Zavoina pseudo-R^2** (McKelvey & Zavoina, 1975) is grounded in the concept of the existence of a latent (unobserved) continuous variable that is theorized to give rise to observed dichotomized values (of 1s and 0s) of a dependent variable. The metric attempts to estimate the proportion of variance explained in the latent variable, as follows:

$$R_{MZ}^2 = \frac{Var(\hat{Y})*(n-1)}{Var(\hat{Y})*(n-1) + n\left(\frac{\pi^2}{3}\right)} \quad (4)$$

where $Var\left(\hat{Y}\right)$ denotes the sample variance of the predicted latent (continuous) values from the logistic model, and $\frac{\pi^2}{3}$ is the theoretical variance of the logistic distribution. Multiple simulation studies (e.g., DeMaris, 2002; Hagle & Mitchell, 1992; Veall & Zimmermann, 1996; Windmeijer, 1995) indicate that this metric most closely approximates the R^2 used in

linear regression. However, computing R^2_{MZ} is more complex than the other metrics discussed and is not routinely provided in popular software packages.

Taken together, the four pseudo-R^2 metrics reviewed will yield quite different results from each other (Smith & McKenna, 2013; Veall & Zimmermann, 1996), and the first two metrics described are simply not good metrics across studies with different sample sizes and baseline probabilities. However, the latter two pseudo-R^2 metrics (R^2_N and R^2_{MZ}) are likely to be quite useful so long as their interpretations are clearly articulated.

What about *Multilevel* Logistic Regression Models?

Nested structures are ubiquitous in social science research where dependencies among and within individuals are likely to exist. In educational research, we collect data on students who are nested within classrooms, classrooms within schools, and schools within districts. In psychological research, individuals may be nested within providers or communities. As well, multiple measurements may be nested within individuals. The defining characteristic of these data structures is the dependency among observations, meaning scores within the same cluster tend to be more like each other than to scores from different clusters (e.g., Goldstein, 2011; Raudenbush, 2002).

Ignoring the nested structure of data can lead to serious statistical problems. First, it violates the fundamental assumption of observation independence in traditional regression models, resulting in distorted standard errors and confidence intervals (e.g., Hox et al., 2010). Second, lower-level predictor effects can be a blend of effects at different levels unless they are decomposed into their sources (e.g., Enders & Tofighi, 2007; Sanders & Konold, 2023). Third, it may lead to ecological fallacy or atomistic fallacy, incorrectly inferring cluster-level relationships to the individual level, or vice versa (Robinson, 1950).

Multilevel logistic regression models can appropriately handle non-independence of observations by accounting for clustering as a random effect (other but less flexible

approaches for handling non-independence exist, such as treating clustering as a multi-categorical predictor fixed effect coupled with cluster-robust standard errors). The simplest two-level random intercept regression model can be expressed as follows.

$$LN\left(\frac{p_{ij}}{1-p_{ij}}\right) = \gamma_{00} + \gamma_{10}X_{1ij} + \gamma_{20}X_{2ij} + \cdots + \gamma_{p0}X_{pij} + \gamma_{01}W_{1j} + \gamma_{02}W_{2j} + \cdots + \gamma_{0q}W_{qj} + u_{0j} \quad (5)$$

In (5), p_{ij} is the probability of success for individual i in cluster j, and the log-odds of p_{ij} is modeled as a function of the sum of the γ_{p0} effects of X_{pij}, which represent individual-level L1 predictor variables (assumed to be cluster-mean centered), the γ_{qj} effects of W_{qj}, which represent cluster-level predictor variables (assumed to be grand-mean centered), and u_{0j}, which is the cluster intercept residual (random effect), assumed to follow a normal distribution with mean 0 and variance τ_{00}.

For multilevel *linear* models, R^2 effect size calculation has been the topic of much research, including most recent work by Rights and Sterba (2019; 2020; 2023) who synthesized earlier work on this topic and then demonstrated the mathematical decomposition of variance for use in a variety of metrics for fixed vs. random effects at each level of a model hierarchy. However, just as with single-level models, these multilevel linear R^2 metrics are not appropriate for multilevel logistic regression models.

Research on Pseudo-R^2 Effect Sizes for Multilevel Logistic Models

Research on effect size measures for multilevel logistic regression models remains notably limited. In their recent systematic review of educational research, Luo et al. (2021) found that, while effect size reporting had increased for multilevel linear models, effect size reporting for multilevel logistic regression models remained scarce. This may be due in part to the relative lack of methodological research on this topic. To date, existing work has primarily proposed isolated metrics without systematic evaluation. For example, Nakagawa et al. (2017) extended the single-level logistic regression McKelvey-Zavoina pseudo-R^2 (R^2_{MZ})

by developing a marginal R^2 (incorporating only fixed effects) and conditional R^2 (incorporating both fixed and random effects) for multilevel logistic regression as follows.

$$R^2_{marginal} = \frac{\sigma_f^2}{\sigma_f^2+\sigma_\alpha^2+\sigma_\varepsilon^2} \quad (6)$$

$$R^2_{conditional} = \frac{\sigma_f^2+\sigma_\alpha^2}{\sigma_f^2+\sigma_\alpha^2+\sigma_\varepsilon^2} \quad (7)$$

In (6) and (7), σ_f^2 represents fixed effects variance, σ_α^2 random effects variance, and σ_ε^2 distribution-specific residual variance = $\frac{\pi^2}{3}$ for logistic regression. While theoretically promising, this approach was not validated using simulations, and the equations given rely on true variances. In practice, sample-based estimates must take sampling error into account as part of the calculation, particularly with respect to cluster size and its interplay with the level of dependency in the data (i.e., the intraclass correlation). As well, the original R^2_{MZ} metric will include a mixture of both fixed and random variation sources (i.e., if the ICC is .50, the metric could display .50 even if the fixed predictors are not predictive). As such, the original metric will be expected to grow as the ICC grows. We therefore propose the following pseudo-R^2 metric based on an adjustment to the original R^2_{MZ} for use with sample-based estimates, with a focus on removing random effects:

$$R^2_{MZ-adj} = \frac{SS_{\hat{Y}}}{SS_{\hat{Y}}+SS_{full\ L2\ residuals}+N(\frac{\pi^2}{3})} * (1-(1-\frac{1}{M})SS_{null\ ICC}) , \quad (8)$$

where $SS_{\hat{Y}}$ is the sum of squared deviations of model-implied predicted values (in logits), $SS_{full\ L2\ residuals}$ is the sum of squared cluster level (L2) residuals (i.e., the model-estimated variance component multiplied by the *df* which is equal to the number of clusters $J - 1$), N is the total sample size at L1, $\frac{\pi^2}{3}$ is the expected variance for the for logistic regression, M is the average cluster size, and $SS_{null\ ICC}$ is the null model (intercept-only) intraclass correlation based on cluster *total* variability $SS_{null\ L2\ residuals}$ (total cluster variability in the model without predictors) divided by the sum of $SS_{null\ L2\ residuals}$ and the total variability for the

logistic regression, $N * \frac{\pi^2}{3}$. If the ICC = 0, then the metric will become the original R^2_{MZ} value. If the ICC > 0, then the original R^2_{MZ} will include the random effects, but the adjustment will down-weight the metric proportionate to the random effects (ICC).

Prevalence of Effect Size Metrics Reported for Multilevel Logistic Models

To assess the current state of effect size reporting in multilevel logistic regression models, a systematic review of published articles between 2020 and 2024 in six major education and psychology journals was undertaken. Journals included those that are prestigious within the field and have high-impact ratings, including: *American Educational Research Journal* (AERJ; 5-year impact factor = 5.1), *Developmental Psychology* (DP; 5-year impact factor = 5.0), *Educational Evaluation and Policy Analysis* (EEPA; 5-year impact factor = 3.7), *Journal of Personality and Social Psychology* (JPSP; 5-year impact factor = 9.2), *Journal of Educational Psychology* (JEP; 5-year impact factor = 6.7), and *Journal of Applied Psychology* (JAP; 5-year impact factor = 11.8). Search terms included: "multilevel logistic regression," "hierarchical logistic regression," and "mixed-effects logistic model." The initial search identified 216 research articles, of which 71 (33%) employed multilevel logistic regression for their analyses.

As shown in Table 1, across the 71 studies, model-level effect size measures were notably absent: only 1 (1%) reported pseudo-R^2 values and 1 (1%) reported marginal or conditional R^2. No studies utilized classification accuracy metrics as effect size measures (i.e., sensitivity, specificity, overall hit rate, or area-under-curve), which was somewhat surprising. Instead, researchers who reported any effect sizes predominantly focused on predictor-specific metrics, with 35 studies (46%) reporting odds ratios, 3 (4%) reporting average marginal effects, and 1 (1%) reporting Cohen's *d*.

As well, there was substantial variation across journals in their reporting practices. JAP stood out with 33% of its multilevel logistic regression studies reporting pseudo-R^2

values, while no such reporting appeared in AERJ, DP, EEPA, JPSP, or JEP. JPSP was the only journal with studies reporting marginal or conditional R^2 (4%). These disparities likely reflect differences in disciplinary traditions and editorial expectations regarding methodological reporting.

It is also notable that, for the few studies that did report effect sizes, methodological transparency was lacking. Researchers rarely provided theoretical justifications for their choice of specific indicators or detailed calculation methods. This systematic gap in effect size reporting not only hinders the interpretability of individual findings but also impedes meaningful cross-study comparisons and knowledge synthesis. The lack of consensus regarding appropriate effect size measures for multilevel logistic regression and insufficient guidance on their interpretation across different research contexts appear to be primary barriers to improved reporting practices in this methodologically complex area.

Current Study

Given the limited research on effect sizes for multilevel logistic regression models and the concerning gaps in reporting practices, this study aims to contribute to the methodological literature by evaluating previous pseudo-R^2 metrics used in single-level logistic models, along with the newly proposed metric in (8), for their performance for 2-level hierarchical logistic regression models. Using Monte Carlo simulation, we evaluated two research questions, as follows.

1) How do traditional pseudo-R^2 measures for single-level models (R^2_{McF}, R^2_{CS}, R^2_N, and R^2_{MZ}) perform in a multilevel logistic regression context?
2) What is the performance of the proposed adjusted McKelvey-Zavoina pseudo-R^2(R^2_{MZ-adj})?

Method

Data Generation

Initial data generation. Data were generated using *Mplus 8* (Muthén & Muthén, 1998/2017) within the *MplusAutomation* package in *R* (Hallquist & Wiley, 2018). We first used a population generating model assuming a continuous latent normally distributed variable (Y^*) with a two-level random-intercept model with three predictors, cluster-mean centered at L1 and grand-mean centered at L2, to reflect a common two-level nested structure in educational and psychological research (e.g., students nested within schools). This model was as follows.

$$Y_{ij} = \gamma_{00} + \gamma_{10}\left(X_{1ij} - \bar{X}_{1.j}\right) + \cdots + \gamma_{30}\left(X_{3ij} - \bar{X}_{3.j}\right) + \gamma_{01}\left(\bar{X}_{1.j} - \bar{X}_{1..}\right) + \cdots + \gamma_{03}\left(\bar{X}_{3.j} - \bar{X}_{3..}\right) + u_{0j} + r_{ij}, \quad (9)$$

In (9) we assumed multivariate normality for the fixed effects and normality for the two random effects. The data-generating parameters were set a priori using a pre-specified correlation matrix among the outcome and predictors to reflect different effect sizes (the correlation structure among predictors was .20 between X_1 and X_2, .10 between X_1 and X_3, and .10 between X_2 and X_3), desired total population R^2 at each level, and the desired intraclass correlation coefficient (ICC). The logistic regression coefficients relating each predictor to $logit(Y)$ were derived as $\beta = R_{XX}^{-1}R_{XY}$, where R_{XX} represents the predictor intercorrelation matrix and R_{XY} represents the predictor-outcome correlations for the continuous logit scale, which were then converted to binary outcomes via threshold cutoffs. For convenience, the ICCs were set equal for the outcome and predictors, which set the L2 variance for the outcome and each predictor to be the same. As a result, the L2 regression coefficients ($\gamma_{01} = \gamma_{03}$) were equivalent in size (respectively) to the L1 regression ($\gamma_{10} = \gamma_{30}$) coefficients.

Conditions that were varied. Four population data-generating conditions were varied to reflect realistic educational psychology research scenarios as well as for consistency with previous multilevel model methodological research, as follows.

1. **Two True Total R^2 Effect Sizes**. To evaluate the performance of pseudo-R^2 metrics under different effect size strengths, the simulation included "weak" and "strong" conditions. The "weak" effect condition simulated a common scenario in actual research where there is a very low amount of true variance explained. In this condition, only X_3 was correlated with Y ($r = .20$) in the data generating model, resulting in a true total $R^2 = .041$ across both levels. The "strong" effect condition simulated stronger predictive relationships, but where each predictor accounts for varying amounts of the true variance but combined, account for a true total $R^2 = .456$ across both levels. Importantly, these R^2 values were built into the data *prior to dichotomization* and were anticipated to be attenuated in the sample data after dichotomization. To summarize, the data-generating model (assuming a linear model) for the "weak" effect size condition was as follows:

$$Y_{ij} = 0 + (-0.012)\left(X_{1ij} - \bar{X}_{1.j}\right) + (-0.012)\left(X_{2ij} - \bar{X}_{2.j}\right) + 0.203\left(X_{3ij} - \bar{X}_{3.j}\right) + (-0.012)\left(\bar{X}_{1.j} - \bar{X}_{1..}\right) + (-0.012)\left(\bar{X}_{1.j} - \bar{X}_{1..}\right) + 0.203\left(\bar{X}_{3.j} - \bar{X}_{3..}\right) + u_{0j} + r_{ij}, \quad (10)$$

and, the true model for the "strong" effect size condition was as follows.

$$Y_{ij} = 0 + 0.282\left(X_{1ij} - \bar{X}_{1.j}\right) + 0.532\left(X_{2ij} - \bar{X}_{2.j}\right) + 0.119\left(X_{3ij} - \bar{X}_{3.j}\right) + 0.282\left(\bar{X}_{1.j} - \bar{X}_{1..}\right) + 0.532\left(\bar{X}_{1.j} - \bar{X}_{1..}\right) + 0.119\left(\bar{X}_{3.j} - \bar{X}_{3..}\right) + u_{0j} + r_{ij}, \quad (11)$$

2. **Three Intraclass Correlation Coefficient Sizes**. The study selected three ICC levels (.10, .20, and .50) were used to represent realistic levels of non-independence in educational research, with .10 being a small clustering effect (10% of variability explained by clustering) and .50 representing strong clustering that would be expected in longer or more intensive intervention studies. The medium ICC value (.20) is typical in educational research (Hedges & Hedberg, 2007). Like the R^2 values, these ICC levels were built into the data

prior to dichotomization and were anticipated to be attenuated in the sample data after dichotomization.

3. **Two Numbers of Clusters (*J*)**. The number of clusters is a critical factor affecting the precision of multilevel model estimates. We selected two representative levels of cluster numbers: 30 clusters (smaller sample) and 100 clusters (larger sample). Thirty clusters represents the minimum acceptable number of clusters commonly found in applied research, while 100 clusters represents a larger and more ideal sample size.
4. **Two Cluster Sizes (*M*)**. Cluster size is equally important as the number of clusters, as it affects the precision of Level 1 estimates and the reliability of ICC estimation. We selected two cluster size levels: 5 observations per cluster (small) and 30 observations per cluster (large). A cluster size of 5 can represent scenarios common in longitudinal studies where measurement occasions form the lower level, while 30 is typical in educational research where students are nested within classrooms.

Fully crossing these conditions produced a total of 2 (effect size magnitude) × 3 (ICC) × 2 (number of clusters) × 2 (cluster size) = 24 simulation conditions. For each condition, 1,000 replicate datasets were generated, totaling 24,000 samples.

Dichotomization of the dependent variable. After the sample data assuming a latent Y^* were generated, two new datasets were created for each of the original data-generating datasets, for a total of 48,000 datasets. Specifically, for each original dataset, the observed values of Y_{ij} were dichotomized to reflect either an idealized balanced scenario ($P(Y_{ij} = 1) = .50$) or the commonly encountered imbalanced situation in real-world data ($P(Y_{ij} = 1) = .20$). Using the known properties of the normal curve in conjunction with relationship between logits and probabilities, we converted the observed Y_{ij} to a Bernoulli Y_{ij}^{bern} so that,

for the balanced case, $Y_{ij}^{bern} = 1\ if\ Y_{ij} \leq 0;\ 0$ otherwise; and for the unbalanced case, $Y_{ij}^{bern} = 1\ if\ Y_{ij} \leq -1.3863;\ 0$ otherwise.

While there is no direct translation of the linear data-generating model regression weights to the logistic multiple regression model (there is no single scaling factor). However, the intercept for the logistic model for the model in which $P(Y_{ij} = 1) = .50$ would be expected to be zero when all Xs = 0, and the intercept for the logistic model in which $P(Y_{ij} = 1) = .20$ have be expected to be -1.39 when all Xs = 0.

Dataset Analyses

Each of the 48,000 simulated datasets were analyzed using the glmer() function in the *R* 'lme4' package (Bates et al., 2015) using two models: a null baseline model (random intercept-only):

$$ln(odds(\Pr(Y_{ij} = 1))) = \gamma_{00} + u_{0j}\,, \tag{12}$$

and a full model including all predictors, reflecting a correctly specified model:

$$ln(odds(\Pr(Y_{ij} = 1))) = \gamma_{00} + \gamma_{10}\left(X_{1ij} - \bar{X}_{1.j}\right) + \cdots + \gamma_{30}\left(X_{3ij} - \bar{X}_{3.j}\right) + \gamma_{01}\left(\bar{X}_{1.j} - \bar{X}_{1..}\right) + \cdots + \gamma_{03}\left(\bar{X}_{3.j} - \bar{X}_{3..}\right) + u_{0j}\,. \tag{13}$$

Both are needed for computing the pseudo-R^2 effect sizes. To handle potential convergence issues, the *bobyqa* maximum likelihood optimizer with a maximum of 100,000 function evaluations was specified (details about nonlinear model estimation optimizers are discussed in Bates et al., 2015).

Simulation Results Analysis Plan

Each of the five previously reviewed pseudo-R^2 metrics (R^2_{McF}, R^2_{CS}, R^2_N, R^2_{MZ}, and R^2_{MZ-adj}) were computed for each dataset according to Eq. 1-4 and 8, respectively, using each dataset's model-based estimates (*R* code for these computations are provided in the Appendix). Model and effect size performance metrics were computed as follows.

Model convergence/extreme values. Any model results with intercept estimates outside the typical range of the normal distribution of log-odds ratios, within ±1.81 standard deviations of the true logit value, were filtered to avoid extreme values that can occur in the sampling process since outliers are not a focus of this study. (The value of 1.81 was chosen because it represents one standard deviation of the logistic distribution by taking the square root of the variance: $\sqrt{\pi^2/3}$.) After filtering, the results sample contained 37,792 (79%) replicate datasets out of the original 48,000. This is described in more detail in the Results.

Model intraclass correlation recovery. Because the intraclass correlation (ICC) estimates are a key part of the newly proposed R^2_{MZ-adj} metric, the extent to which the model recovered ICC values was computed for each null model as $\frac{\hat{\sigma}^2_{between-cluster}}{\hat{\sigma}^2_{between-cluster}+\left(\frac{\pi^2}{3}\right)}$, where $\hat{\sigma}^2_{between-cluster}$ is the model-estimated random intercept variance (in logits) and $\frac{\pi^2}{3}$ is the assumed within-cluster variance of the standard logistic distribution.

Model sensitivity prediction accuracy. Model sensitivity accuracy for the full model with predictors was computed for each dataset by comparing the model-predicted probabilities to the observed binary outcome values. A threshold predicted probability of .50 was used to indicate a 1, otherwise 0.

Pseudo-R^2 effect size raw bias. Raw bias was computed as $R^2_{est} - R^2_{true}$, where R^2_{est} is the estimated pseudo-R^2 value from the model and R^2_{true} is the true total R^2 value used in the original data generation process (i.e., based on the linear model).

Results

Although not the focus of this paper, the logistic regression model coefficient estimates across conditions are given in Table 3. Thereafter, Tables 4 and Table 5 display cell mean results for the balanced ($P(Y = 1) = .50$) condition and imbalanced simulation condition ($P(Y = 1) = .20$), respectively, by each performance metric, including convergence (filtering

for extreme intercept values), ICC recovery, prediction sensitivity, and raw bias in the pseudo-R^2 metrics.

Model Convergence

Model convergence rates were strongly influenced by the baseline probability (i.e., balance) of the binary outcome variable. Models with balanced outcomes (mean probability of 50%) achieved convergence rates near 100% across all conditions. In contrast, models with imbalanced outcomes (mean probability of 20%) exhibited substantially lower convergence rates, particularly under lower ICC conditions typical in educational research settings (see Table 2; Figure 1 displays results for the strong effect size condition). This suggests that researchers working with very rare or very common events should always check their results for extreme values and consider removing potentially problematic predictors.

Model ICC Recovery

Model-estimated ICC values consistently underestimated the data-generating ICCs by approximately 5-10% (Tables 1-2; Figure 2). This underestimation was slightly less pronounced with larger cluster sizes ($M = 30$ vs. $M = 5$) and more clusters ($J = 100$ vs. $J = 30$). This attenuation aligns with theoretical expectations regarding the dichotomization of continuous variables, which typically reduces observed correlations (e.g., Hunter & Schmidt, 1990).

Model Sensitivity Accuracy

The model sensitivity accuracy rates (correctly classified 1 values) varied considerably across conditions but generally increased as the ICC increased (Tables 1-2; Figure 3 displays the pattern for the strong effect size conditions). The balanced condition (50% prevalence) achieved approximately 80% accuracy, whereas the imbalanced condition (20% prevalence) only reached about 40% accuracy. In both cases, prediction accuracy improved with increasing ICC values, with this effect being particularly pronounced in the

imbalanced condition. While the 95% confidence intervals for different sample size combinations overlapped considerably, the imbalanced condition showed slightly better prediction accuracy wither cluster sizes (M = 30 vs. M = 5).

Pseudo-R^2 Raw Bias

The comparison of pseudo-R^2 metrics revealed substantial differences in accuracy and consistency across simulation conditions. Specifically, for the weak (near zero) effect size conditions, the traditional metrics were close to zero but the newer metrics were somewhat inflated (Tables 1-2). For the strong effect size conditions, the adjusted McKelvey-Zavoina R^2 was the least biased (Figure 4) whereas the more traditional metrics consistently underestimated the true effect size and the un-adjusted McKelvey-Zavoina R^2 exhibited overestimation as the ICC increased.

Discussion

The present study contributes to the literature on effect size reporting for multilevel logistic regression models in two ways: first, a systematic review of applied research using logistic regression in top-tier education and psychology journals revealed that researchers are largely not reporting any model-based effect sizes other than regressor odds ratios. Second, our simulation study revealed that, in strong effect size conditions, the adjusted McKelvey-Zavoina R^2 metric may be a promising novel approach for 2-level hierarchical logistic regression models when compared to traditional pseudo-R^2 metrics that underestimate the true R^2 as well as the more recent Nakagawa et al. (2017) McKelvey-Zavoina R^2 metric that over-estimates the true R^2 affected by increased ICC levels. This said, in near-null conditions the new metric overestimates the true R^2 similar to the un-adjusted McKelvey-Zavoina R^2.

In terms of ICC recovery and model sensitivity prediction, our findings are consistent with literature on the attenuation effect of dichotomization on correlations (Hunter & Schmidt, 1990) as well as the impact of imbalanced data on model performance (Hemmert et

al., 2018). In particular, the significantly reduced model convergence rates and prediction sensitivity under imbalanced binary outcomes have important implications for disciplines studying rare events (such as special education or epidemiology): these datasets may need to consider dropping certain predictors or using other modeling strategies such as generalized estimating equations (McNeish et al., 2017).

Limitations and Future Research

The present study has several limitations that can be addressed in future work. First, the simulations were limited to two-level random intercept models; future research should extend to random slopes models, which are common in applied research. The random effects structure in multilevel models may have complex impacts on effect size estimation that require further investigation. Second, this study did not examine coverage rates (and variability in the estimates) which should be part of future work. Third, this study focused only on whole-model-level pseudo-R^2 metrics and did not address potential level-based predictor effect size metrics analogous to squared semi-partial correlations (sr^2) that may help to avoid reliance on odds ratios that are difficult to interpret and rely on intercept probability rates. Fourth, a Pearsonian approach was used to tabulate pseudo-R^2 bias estimation (i.e., based on the notion of a latent Y^* underlying the observed binary Y values); there may be other approaches to explore this idea in the future.

Additionally, if the adjusted McKelvey-Zavoina pseudo-R^2 metric can be further studied to reduce bias in near-zero conditions, then a standardized interpretation framework will need to be established. While Cohen (1988) proposed thresholds for small, medium, and large effects (.01, .09, .25) for linear model R^2, there is no evidence that these thresholds will apply to the pseudo-R^2 metrics as they stand. As well, extending effect size explorations to Bayesian estimation as well as ordinal or multinomial logistic regression models (i.e., other types of categorical dependent variables) can also serve as exciting future directions. Finally,

any promotion of pseudo-R^2 metrics will require a user-friendly wrap-around package for *R* lme4 and nlme packages to be implemented. The practical value of effect size metrics depends not only on their statistical properties but also on researchers' ability to easily calculate and correctly interpret them.

Conclusion

Reporting effect sizes is an important staple of modern research as they facilitate the quantification of variable relations for meta-analyses irrespective of operational scales used, enable policymakers to understand practical returns on a program investment, help alleviate reliance on null hypothesis testing, and are an essential ingredient for planning research sample sizes. In single-level logistic regression, it is well-known that traditional variance-explained metrics like those in linear regression are not possible; instead, pseudo-R^2 measures that capture how much better a model with regressors fits compared to a null model are in wide use and are available in popular statistical computing packages. The present study shows that most research employing multilevel logistic models do not report any effect sizes, likely due to a lack of research on the topic. Our simulations here show the potential promise of a new pseudo-R^2 metric for 2-level hierarchical logistic regression models that can be used as a starting point for additional research to better facilitate study effect size reporting.

References

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences (2nd ed.).* Hillsdale, NJ: Lawrence Erlbaum Associates, Publishers.

Cox, D. R., & Snell, E. J. (1989). *Analysis of Binary Data* (2nd ed.). Chapman and Hall.

DeMaris, A. (2002). Explained variance in logistic regression: A Monte Carlo study of proposed measures. *Sociological Methods & Research*, *31*(1), 27–74. https://doi.org/10.1177/0049124102031001002

Enders, C. K., & Tofighi, D. (2007). Centering predictor variables in cross-sectional multilevel models: A new look at an old issue. *Psychological Methods*, *12*(2), 121. https://doi.org/10.1037/1082-989X.12.2.121

Fritz, C. O., Morris, P. E., & Richler, J. J. (2012). Effect size estimates: Current use, calculations, and interpretation. *Journal of Experimental Psychology: General*, *141*(1), 2-18. https://doi.org/10.1037/a0024338

Goldstein, H. (2011). *Multilevel Statistical Models*. John Wiley & Sons.

Hagle, T. M., & Mitchell, G. E. (1992). Goodness-of-fit measures for probit and logit. *American Journal of Political Science*, 762–784. https://doi.org/10.2307/2111590

Hallquist, M. N., & Wiley, J. F. (2018). MplusAutomation: An R package for facilitating large-scale latent variable analyses in Mplus. *Structural Equation Modeling: A Multidisciplinary Journal, 25*(4), 621–638. https://doi.org/10.1080/10705511.2017.1402334

Hemmert, G. A. J., Schons, L. M., Wieseke, J., & Schimmelpfennig, H. (2018). Log-likelihood-based pseudo-R^2 in logistic regression: Deriving sample-sensitive

benchmarks. *Sociological Methods & Research*, *47*(3), 507–531. https://doi.org/10.1177/0049124116638107

Hoogland, J. J., & Boomsma, A. (1998). Robustness studies in covariance structure modeling: An overview and a meta-analysis. *Sociological Methods & Research, 26*(3), 329–367. https://doi.org/10.1177/0049124198026003003

Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2000). *Applied Logistic Regression.* Hoboken. NJ: John Wiley & Sons.

Hox, J. J., Maas, C. J. M., & Brinkhuis, M. J. S. (2010). The effect of estimation method and sample size in multilevel structural equation modeling. *Statistica Neerlandica*, *64*(2), 157–170. https://doi.org/10.1111/j.1467-9574.2009.00445.x

Hunter, J. E., & Schmidt, F. L. (1990). Dichotomization of continuous variables: The implications for meta-analysis. *Journal of Applied Psychology, 75*(3), 334–349. https://doi.org/10.1037/0021-9010.75.3.334

Long, J. S. (1997). *Regression Models for Categorical and Limited Dependent Variables.* Sage Publications.

McFadden, D. (1974). Conditional logit analysis of qualitative choice behaviour. In P. Zarembka (Ed.), *Frontiers in Econometrics* (pp. 105–142). Academic Press.

McFadden, D. (1979). Quantitative methods for analyzing travel behaviour of individuals: Some recent developments. In D. A. Hensher & P. R. Stopher (Eds.), *Behavioural Travel Modelling* (pp. 279-318). Croom Helm.

McKelvey, R. D., & Zavoina, W. (1975). A statistical model for the analysis of ordinal level dependent variables. *The Journal of Mathematical Sociology*, *4*(1), 103–120. https://doi.org/10.1080/0022250X.1975.9989847

McNeish, D., Stapleton, L. M., & Silverman, R. D. (2017). On the unnecessary ubiquity of hierarchical linear modeling. *Psychological Methods, 22*(1), 114–140. https://doi.org/10.1037/met0000078

Menard, S. (2000). Coefficients of determination for multiple logistic regression analysis. *The American Statistician*, *54*(1), 17–24. https://doi.org/10.1080/00031305.2000.10474502

Muthén, L.K. and Muthén, B.O. (1998-2017). *Mplus User's Guide*, *8th Ed*. Los Angeles, CA: Muthén & Muthén.

Nagelkerke, N. J. (1991). A note on a general definition of the coefficient of determination. *Biometrika, 78*(3), 691-692. https://doi.org/10.1093/biomet/78.3.691

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical Linear Models*. Thousand Oaks, CA: Sage.

Rights, J. D., & Sterba, S. K. (2019). Quantifying explained variance in multilevel models: An integrative framework for defining *R*-squared measures. *Psychological Methods, 24*(3), 309–338. https://doi.org/10.1037/met0000184

Rights, J. D., & Sterba, S. K. (2020). New recommendations on the use of *R*-squared differences in multilevel model comparisons. *Multivariate Behavioral Research, 55*(4), 568–599. https://doi.org/10.1080/00273171.2019.1660605

Rights, J. D., & Sterba, S. K. (2023). *R*-squared measures for multilevel models with three or more levels. *Multivariate Behavioral Research, 58*(2), 340–367. https://doi.org/10.1080/00273171.2021.1985948

Robinson, W. S. (1950). Ecological correlations and the behavior of individuals. *American Sociological Review*, *15*(3), 351–357. https://doi.org/10.2307/2087176

Sanders, E. A., & Konold, T. R. (2023). X matters too: How the blended slope problem manifests differently in unilevel vs. multilevel models. *Methodology, 19*(1), 1-23. https://doi.org/10.5964/meth.9925

Smith, T. J., & McKenna, C. M. (2013). A comparison of logistic regression pseudo R^2 indices. *Multiple Linear Regression Viewpoints*, *39*(2), 17–26.

Thompson, B. (2007). Effect sizes, confidence intervals, and confidence intervals for effect sizes. *Psychology in the Schools*, *44*(5), 423–432. https://doi.org/10.1002/pits.20234

Veall, M. R., & Zimmermann, K. F. (1996). Pseudo-R^2 measures for some common limited dependent variable models. *Journal of Economic Surveys*, *10*(3), 241–259. https://doi.org/10.1111/j.1467-6419.1996.tb00013.x

Wilkinson, L. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*, *54*(8), 594-604. https://doi.org/10.1037/0003-066X.54.8.594

Windmeijer, F. A. G. (1995). Goodness-of-fit measures in binary choice models. *Econometric Reviews*, *14*(1), 101–116. https://doi.org/10.1080/07474939508800306

Table 1

Systematic Review Results for Multilevel Logistic Model Effect Size Reporting in Educational and Psychological Research, 2020-2024

Journal	Total Research Articles	Multilevel Logistic Model Used		*Total ES*						*Coefficient ES*					
				Pseudo-R^2		Marginal or Conditional R^2		Classification Accuracy		Odds Ratio		Average Marginal Effects		Cohen's *d*	
		N	(%)	N	(%)	N	(%)	N	(%)	N	(%)	N	(%)	N	(%)
AERJ	59	9	15%	0	0%	0	0%	0	0	3	33%	3	33%	0	0%
DP	39	26	67%	0	0%	0	0%	0	0	13	46%	0	0%	0	0%
EEPA	49	2	4%	0	0%	0	0%	0	0	1	50%	0	0%	0	0%
JPSP	44	23	52%	0	0%	1	4%	0	0	11	46%	0	0%	1	1%
JEP	19	8	42%	0	0%	0	0%	0	0	4	50%	0	0%	0	0%
JAP	6	3	50%	1	33%	0	0%	0	0	3	60%	0	0%	0	0%
Total	**216**	**71**	**33%**	**1**	**1%**	**1**	**1%**	**0**	**0%**	**35**	**46%**	**3**	**4%**	**1**	**1%**

Note. ES = Effect Size; AERJ = *American Educational Research Journal*; DP = *Developmental Psychology*; EEPA = *Education Evaluation and Policy Analysis*; JPSP = *Journal of Personality and Social Psychology*; JEP = *Journal of Educational Psychology*; JAP = *Journal of Applied Psychology*.

Table 2

Simulation Design Conditions

Condition	No. Levels	Levels
Intraclass correlation	3	.10, .20, .50
Total effect size	2	Weak true effect size R^2 = .041 Strong true effect size R^2 = .456
J Clusters	2	30, 100
M Cluster Size	2	5, 30
Binary Outcome Balance	2	Balanced (.50), Unbalanced (.20)

Note. Design conditions fully crossed for a total of 24 cells prior to dichotomization of the dependent variable (48 cells after dichotomization), with 1,000 replications drawn for each cell. Each dataset was analyzed using a "null" intercept-only model and a "full" model with three L1 and three L2 predictors.

Table 3

Simulation Results Coefficient Estimates (in Logits)

Coefficient		Weak R^2		Strong R^2	
		Y_p = .20	Y_p = .50	Y_p = .20	Y_p = .50
Intercept		-2.51	0.00	-2.97	0.00
L1	X1	0.04	0.03	-0.66	-0.68
	X2	0.03	0.03	-1.21	-1.27
	X3	-0.41	-0.36	-0.29	-0.29
L2	X1	0.04	0.03	-0.65	-0.67
	X2	0.03	0.03	-1.24	-1.27
	X3	-0.39	-0.35	-0.27	-0.28

Table 4

Simulation Performance by Design Condition for the Balanced Outcome with Pr(Y=1) = .50

True R^2		J	M	True ICC	Converg Rate	Est ICC	Est Sens	Raw Bias				
								McFadden R^2_{McF}	Cox-Snell R^2_{CS}	Nagel R^2_N	M-Z R^2_{MZ}	Adj M-Z R^2_{MZ-adj}
.04	(weak)	30	5	.1	1.00	.06	.65	.01	.03	.05	.07	.07
				.2	1.00	.13	.69	.02	.03	.05	.10	.10
				.5	1.00	.41	.80	.02	.03	.05	.22	.18
			30	.1	1.00	.07	.63	-.01	.00	.01	.05	.05
				.2	1.00	.14	.67	-.01	-.01	.01	.11	.10
				.5	1.00	.43	.76	-.01	-.01	.00	.27	.21
		100	5	.1	1.00	.06	.65	-.01	.00	.01	.03	.03
				.2	1.00	.14	.71	-.01	.00	.01	.06	.06
				.5	.99	.43	.80	-.01	.00	.01	.19	.16
			30	.1	1.00	.08	.64	-.02	-.01	.00	.04	.04
				.2	1.00	.16	.67	-.02	-.01	.00	.11	.10
				.5	1.00	.46	.76	-.02	-.01	.00	.27	.20
.46	(strong)	30	5	.1	1.00	.07	.77	-.13	-.09	.02	.07	.07
				.2	1.00	.14	.79	-.13	-.10	.01	.09	.07
				.5	.99	.43	.86	-.13	-.12	.00	.14	.05
			30	.1	1.00	.07	.76	-.16	-.12	-.02	.04	.04
				.2	1.00	.16	.78	-.16	-.13	-.03	.07	.06
				.5	1.00	.44	.83	-.16	-.17	-.05	.15	.04
		100	5	.1	1.00	.07	.77	-.15	-.11	-.01	.04	.03
				.2	1.00	.16	.79	-.15	-.12	-.01	.05	.04
				.5	1.00	.45	.86	-.16	-.14	-.03	.11	.02
			30	.1	1.00	.08	.76	-.16	-.12	-.02	.04	.03
				.2	1.00	.17	.78	-.16	-.13	-.03	.07	.05
				.5	1.00	.47	.83	-.16	-.18	-.06	.14	.02

Note. Each cell originally had 1,000 replications. Converg = convergence represented as model-based estimates that did not exhibit extreme values. Mean ICC, sensitivity, and pseudo-R^2 values are based on estimates that were filtered for extreme values.

Table 5

Simulation Performance by Design Condition for the Balanced Outcome with Pr(Y=1) = .20

True R^2		J	M	True ICC	Converg Rate	Est ICC	Est Sens	Raw Bias				
								McFadden R^2_{McF}	Cox-Snell R^2_{CS}	Nagel R^2_N	M-Z R^2_{MZ}	Adj M-Z R^2_{MZ-adj}
.04	(weak)	30	5	.1	.79	.06	.03	.04	.01	.07	.14	.14
				.2	.79	.12	.06	.04	.02	.07	.15	.14
				.5	.77	.37	.32	.03	.02	.06	.21	.18
			30	.1	1.00	.08	.00	.00	-.02	.01	.07	.07
				.2	1.00	.17	.03	.00	-.02	.01	.12	.12
				.5	.92	.44	.34	-.01	-.02	.00	.26	.20
		100	5	.1	.97	.10	.01	.00	-.01	.02	.07	.07
				.2	.97	.19	.02	.00	-.01	.02	.09	.08
				.5	.92	.46	.33	.00	-.01	.01	.18	.14
			30	.1	1.00	.10	.00	-.01	-.02	.00	.06	.06
				.2	1.00	.20	.03	-.01	-.02	-.01	.11	.11
				.5	.99	.49	.35	-.01	-.02	-.01	.25	.18
.46	(strong)	30	5	.1	.17	.03	.26	-.15	-.24	-.08	.05	.04
				.2	.23	.07	.30	-.14	-.22	-.06	.07	.07
				.5	.33	.29	.55	-.13	-.19	-.04	.12	.07
			30	.1	.19	.06	.22	-.16	-.25	-.09	.04	.04
				.2	.26	.14	.32	-.16	-.24	-.08	.07	.06
				.5	.35	.42	.59	-.16	-.23	-.08	.14	.04
		100	5	.1	.17	.03	.20	-.16	-.25	-.09	.04	.04
				.2	.22	.08	.27	-.15	-.24	-.08	.05	.05
				.5	.30	.39	.55	-.16	-.22	-.08	.09	.02
			30	.1	.07	.08	.23	-.16	-.26	-.09	.05	.04
				.2	.14	.18	.32	-.16	-.25	-.09	.07	.05
				.5	.25	.46	.59	-.16	-.24	-.09	.13	.01

Note. Each cell originally had 1,000 replications. Converg = convergence represented as model-based estimates that did not exhibit extreme values. Mean ICC, sensitivity, and pseudo-R^2 values are based on estimates that were filtered for extreme values.

Figure 1

Convergence Rates of Multilevel Logistic Regression Models for Strong R^2 Conditions

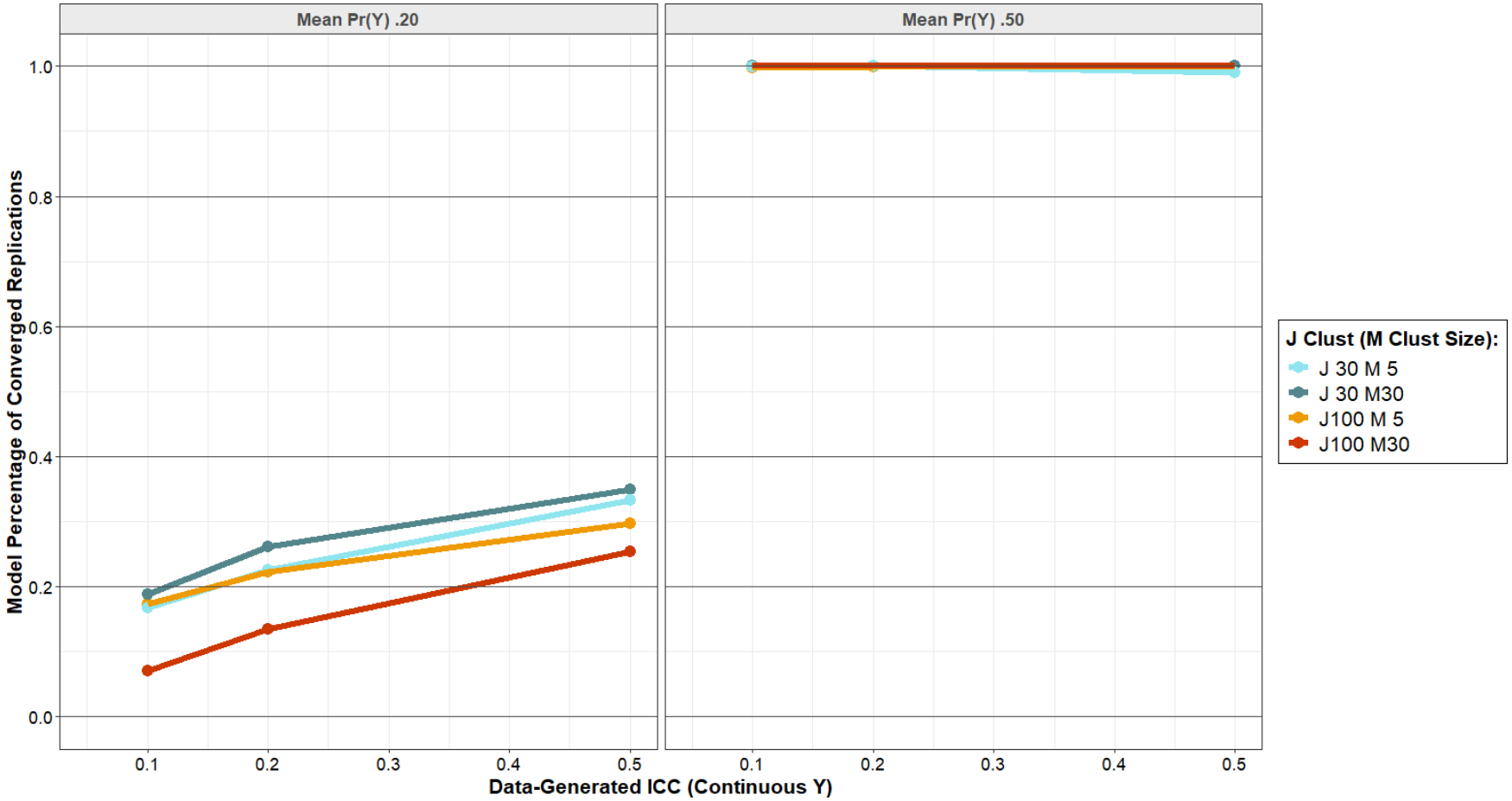


Figure 2

Comparison of Model-Estimated ICC Estimates and True ICC Values for Strong R^2 Conditions

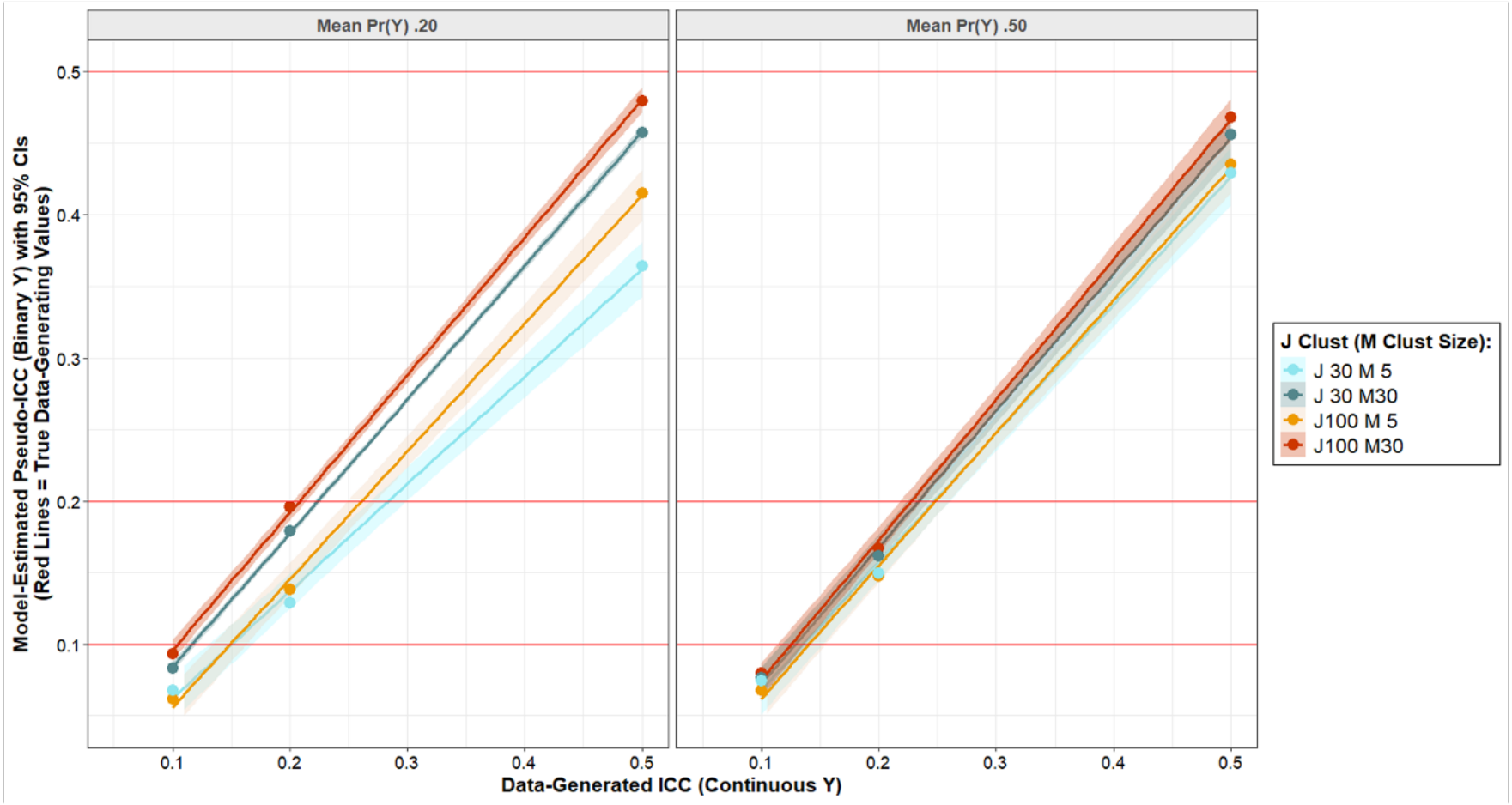


Figure 3

Prediction Sensitivity of Multilevel Logistic Regression Models Across Conditions for Strong R^2 Conditions

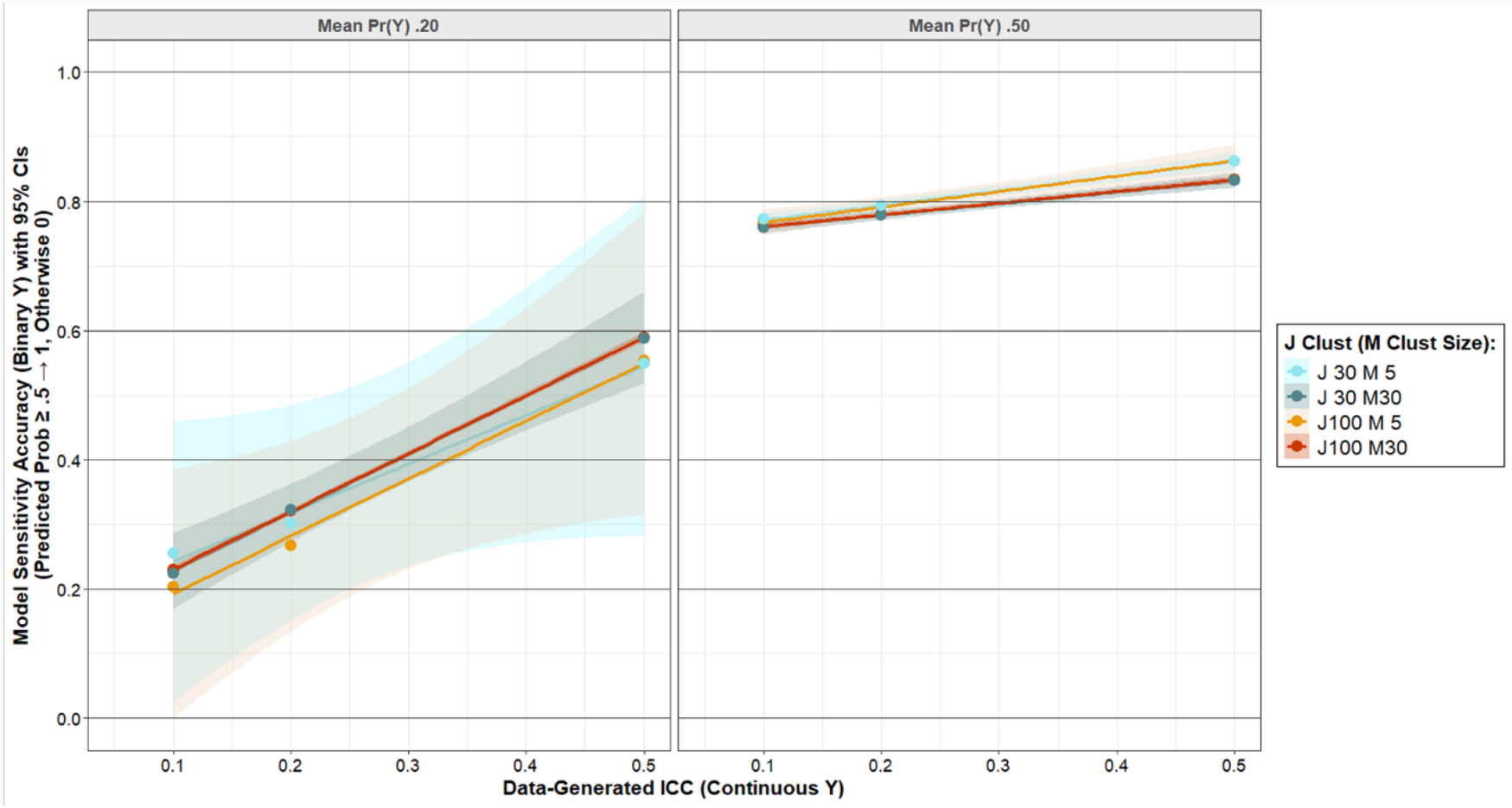
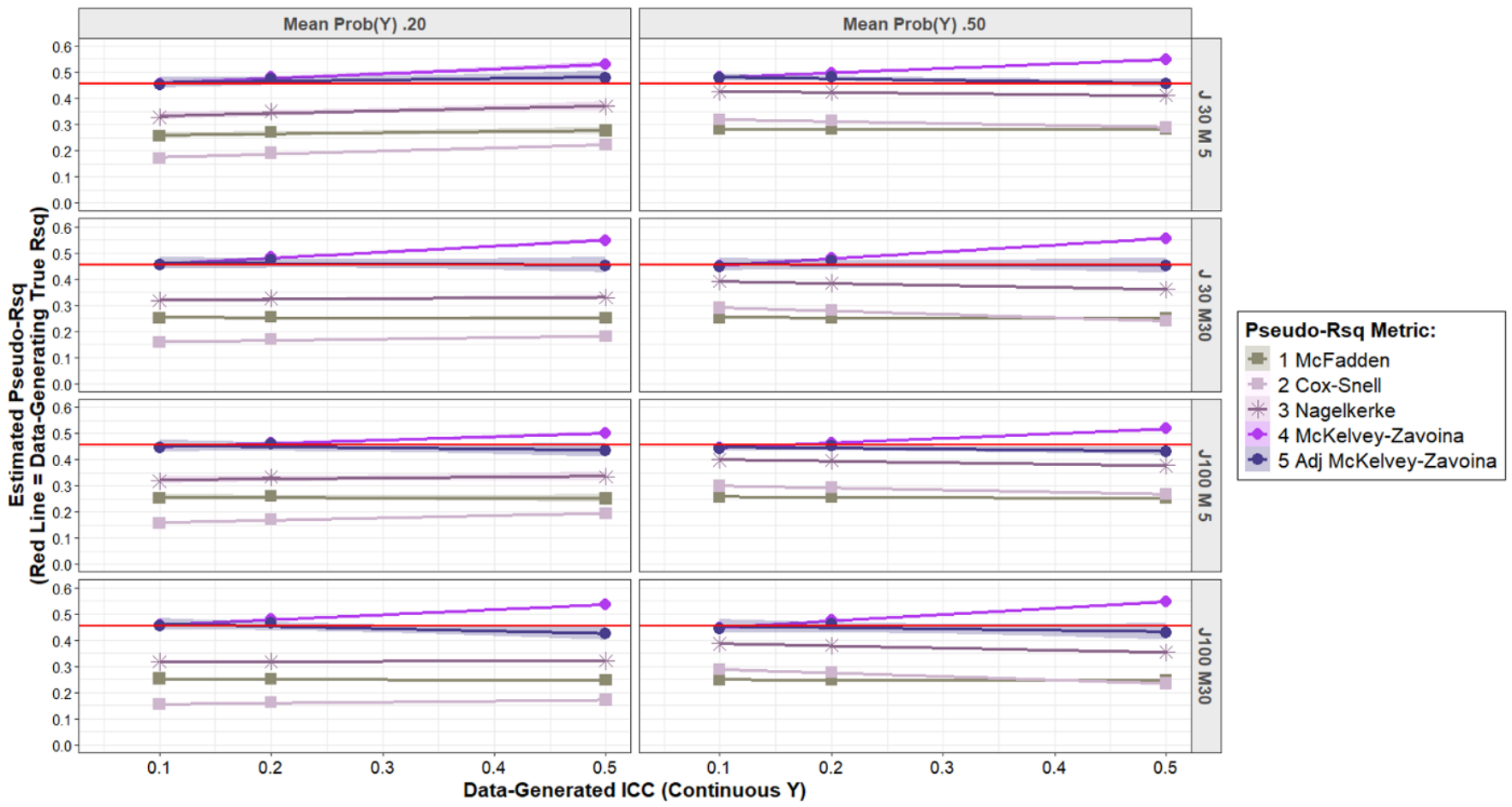


Figure 4

Estimated vs. True Effect Size for different Pseudo-R^2 Metrics for Strong R^2 Conditions



Appendix

Code for Pseudo-R^2 Calculations for Model-Based Estimates

```
# Required Packages
library(dplyr)
library(car)
library(tidyr)

# Function: Calculate Pseudo R² Measures
calculate_pseudo_r2 <- function(null_model_data, full_model_data) {

  # McFadden R²
  mcfadden_r2 <- 1 - (full_model_data$LogLik / null_model_data$LogLik)

  # Cox-Snell R²
  n <- full_model_data$N
  cox_snell_r2 <- 1 - exp((2/n) * (null_model_data$LogLik - full_model_data$LogLik))

  # Nagelkerke R²
  nagelkerke_r2 <- cox_snell_r2 / (1 - exp(-null_model_data$Deviance/n))

  # McKelvey-Zavoina R²
  # Calculate sum of squares from fitted values (matching your original code)
  SS_null <- null_model_data$null_fittedvar * (n - 1)
  SS_full <- full_model_data$full_fittedvar * (n - 1)

  # Basic MZ R² calculations
  mz_r2_null <- SS_null / (SS_null + n * null_model_data$L2var + n * (pi^2)/3)
  mz_r2_full <- SS_full / (SS_full + n * full_model_data$L2var + n * (pi^2)/3)

  return(list(
    mcfadden = mcfadden_r2,
    cox_snell = cox_snell_r2,
    nagelkerke = nagelkerke_r2,
    mckelvey_zavoina_null = mz_r2_null,
    mckelvey_zavoina_full = mz_r2_full
  ))
}

# Function: Calculate Adjusted McKelvey-Zavoina R²
calculate_adjusted_mz <- function(mz_r2_full, l2var_full, sample_size, cluster_size) {
  # Calculate ICC using SS method
  icc_full_ss_est <- (l2var_full * (sample_size - 1)) /
                     (l2var_full * (sample_size - 1) + sample_size * (pi^2)/3)

  # Apply adjustment formula from your original code
  adjustment <- 1 - (1 - 1/cluster_size) * (icc_full_ss_est^2)
  adjusted_mz <- mz_r2_full * adjustment

  return(adjusted_mz)
}
```