sociological science

Inequality and Total Effect Summary Measures for Nominal and Ordinal Variables

Trenton D. Mize, Bing Han

Purdue University

Abstract: Many of the topics most central to the social sciences involve nominal groupings or ordinal rankings. There are many cases in which a summary of a nominal or ordinal independent variable's effect, or the effect on a nominal or ordinal outcome, is needed and useful for interpretation. For example, for nominal or ordinal independent variables, a single summary measure is useful to compare the effect sizes of different variables in a single model or across multiple models, as with mediation. For nominal or ordinal dependent variables, there are often an overwhelming number of effects to examine and understanding the holistic effect of an independent variable or how effect sizes compare within or across models is difficult. In this project, we propose two new summary measures using marginal effects (MEs). For nominal and ordinal independent variables, we propose *ME inequality* as a summary measure of a nominal or ordinal independent variable's holistic effect. For nominal and ordinal outcome models, we propose a *total ME* measure that quantifies the comprehensive effect of an independent variable across all outcome categories. The added benefits of our methods are both intuitive and substantively meaningful effect size metrics and approaches that can be applied across a wide range of models, including linear, nonlinear, categorical, multilevel, longitudinal, and more.

Keywords: nominal variables; ordinal variables; categorical data analysis; marginal effects; inequality; statistics

Reproducibility Package: All data and coding files needed to reproduce all results shown in this article are available at both `www.trentonmize.com/research` and OSF (`osf.io/myehf/`). In addition to the replication files, simplified template/example Stata and R files are also available in the same locations.

Citation: Mize, Trenton D., Bing Han. 2025. "Inequality and Total Effect Summary Measures for Nominal and Ordinal Variables" Sociological Science 12: 115-157.

Received: November 27, 2024

Accepted: January 7, 2025

Published: February 5, 2025

Editor(s): Arnout van de Rijt, Kristian B. Karlson

DOI: 10.15195/v12.a7

Many variables of fundamental interest to social scientists are nominal groupings or ordinal rankings. A few examples are country of birth, educational attainment, political party, religious affiliation, social class, race-ethnicity, relationship status, and occupational sector. There are many cases in which a summary of a nominal or ordinal independent variable's (IV) effect as a holistic construct, or the comprehensive effect on a nominal or ordinal outcome, is needed and useful for interpretation. For example, does educational attainment or occupational sector have a larger impact on wages? Both variables in this case involve multiple categories and answering this question requires somehow quantifying each IVs' holistic effect. As another example, how much of the racial-ethnic disparities in health are explained by socioeconomic status (SES)? Here, summarizing the effect of race-ethnicity before and after accounting for SES factors is needed along with a test of the difference in the summaries. As a last example, does age or educational attainment have a larger impact on self-rated health? Here, self-rated health is usually measured in four or five categories, and understanding effects requires examining the effect of age and education on each outcome category separately making it difficult to provide a single answer to the question.

In this article, we propose two new summary measures for nominal and ordinal variables. For nominal and ordinal IVs, we build on multiple classic approaches that remain useful but limited in their application and/or interpretation. We propose a measure of inequality using marginal effects (MEs) to summarize a nominal or ordinal IVs' holistic effect. The benefits of our method are both an intuitive and substantively meaningful effect size metric and an approach that can be applied across many types of models. Existing methods are mostly limited to linear models and/or provide an effect size that does not clearly represent the substantive significance of the IV.

In brief, our *ME inequality* method involves calculating pairwise comparisons between the predictions for each category of a nominal or ordinal IV and then averaging the comparisons for a measure of mean inequality. This provides a single number and confidence interval for the holistic effect of a nominal or ordinal IV. For example, our method can provide answers to questions such as "Does race-ethnicity matter for this outcome?" and "If so, how much does it matter, on average?" Furthermore, we provide two versions of the *ME inequality* measure that does or does not account for differences in the sizes of the groups/categories of the variable and discuss benefits and potential applications of each approach.

Nominal and ordinal dependent variables propose many similar challenges for interpretation. There are at least as many effects for each IV as there are outcome categories, making a holistic understanding difficult. We develop a *total ME* measure that quantifies the holistic effect of an IV across all outcome categories. In brief, for continuous and binary IVs, we sum the absolute values of the MEs on all outcome categories for a summary measure. In cases where both the independent and dependent variables are nominal or ordinal, we use our new *ME inequality* measures in the calculation of the *total ME*.

To begin, we outline how nominal and ordinal IVs are traditionally analyzed in statistical models along with some recent criticisms of this approach. We next discuss existing summary measures of nominal and ordinal IVs effects. We then develop our measures of *ME inequality* using the marginal effects framework for interpretation and illustrate our approach in a series of applied examples. Finally, we discuss nominal and ordinal outcome models and the difficulty of summarizing effects in these models. We then develop new *total ME* summary measures of effects in these models and highlight their utility in several examples.

We include example Stata and R files that recreate all the examples shown in this article with annotated and simplified code; we intend for these to be the primary files used by applied researchers to understand implementing our approach. These, along with more complex replication files, can be downloaded at https://www.trentonmize.com/research.

Nominal and Ordinal IVs

Nominal and Ordinal Measurement Levels

Nominal variables are those with two or more categories that cannot be ordered on a single dimension (Stevens 1946; Long 1997). Number values are assigned to the categories, though the numbers and their ordering are arbitrary (e.g., 1 = Black, 2 = Asian, 3 = White, 4 = Hispanic, etc.). This arbitrary nature of the numbering makes most descriptive statistics such as the mean meaningless (Melamed and Doan 2023). Instead, we usually descriptively examine a nominal variable by calculating the proportion/percentage of observations in each category (e.g., the U.S. population as of 2024 is 59 percent White non-Hispanic, 14 percent Black, and 19 percent Hispanic). Note that binary variables are nominal variables with only two categories. But usually, the term "nominal" is used to refer specifically to variables with three or more categories though everything we cover in this article applies equally to the binary case.

Ordinal variables similarly have finite categories and no true continuous metric. However, the ordering of the categories is meaningful with only a single underlying dimension affecting the ordering being necessary to strictly qualify as ordinal (Stevens 1946; Long 1997). For example, educational attainment measured in degrees is a prototypical case (e.g., 1 = no high school degree, 2 = high school degree, 3 = college degree, and 4 = graduate degree). Here, the numbers carry information on the ordering but not for the distance between categories, that is, the spacing implied by the numbers assigned to the categories is arbitrary and no meaning is implied by it.

Traditional Nominal and Ordinal IV Interpretations in Statistical Models

An ordinal specification is not an option for an IV. In what follows, we do not delineate between nominal and ordinal IVs because there is no distinction in the regression modeling framework. In a regression model, there are only two options for how to specify an IV: continuous or nominal (where binary is a special case of nominal). Ordinal variables are a common case where analysts must decide between these two imperfect choices. However, continuous variables can also sometimes be specified as nominal in a model; for example, although age is a continuous variable, a nominal specification of age as generations could be appropriate.

For the ordinal IV case, consider a four-category Likert style variable where the categories are 1 = strongly disagree, 2 = disagree, 3 = agree, and 4 = strongly agree. Is it appropriate to treat this variable as a continuous IV in a model? Doing so would not only impose that the ordering of the variables is indeed exactly that implied by the number values (potentially reasonable) but also that the spacing between the categories is equal. For example, that the gap between disagree and agree represents the same difference in absolute agreement as the gap between agree and strongly agree (usually unlikely). Another potential issue is that variables such as Likert style questions are ordered on more than one dimension, in this case agreement (as intended) and also intensity of opinion: those who strongly disagree and those who strongly agree are similar in intensity of opinion (Long and Freese 2014). An easy and oft-used solution is to treat the ordinal variable as nominal in the model, which will not impose ordering or equidistant assumptions and restrictions. In what follows, the methods for nominal IVs outlined apply equally to ordinal IVs that are specified as nominal in the model.

Nominal IV specifications and interpretations. For nominal and ordinal IVs, the traditional approach to statistical modeling (e.g., a regression model) is to specify a series of binary indicators for each category of the variable and omit one category as a reference. Each coefficient then represents the difference between a specific category and the reference category. For example, consider racial-ethnic gaps in average earnings in the United States. If our data include a four-category nominal *race-ethnicity* variable and a continuous *hourly wages* variable, we could fit a linear regression model omitting White as the reference category:

$$\widehat{wages_i} = \hat{\beta}_0 + \hat{\beta}_{Black}\, Black_i + \hat{\beta}_{Hispanic}\, Hispanic_i + \hat{\beta}_{other}\, other_i \ldots. \tag{1}$$

We include controls for *age* and *gender* but don't show their coefficients here. For this and most examples in this article, we use the General Social Survey (GSS), which is a nationally representative survey of the adult U.S. population. Using only the 2021 GSS data, we obtain estimates:

$$\widehat{wages_i} = 14.729 - 3.440\, Black_i - 3.766\, Hispanic_i + 6.664\, other_i \ldots. \tag{2}$$

All coefficients in the model are significant at the $p < 0.05$ level in a two-tailed test. Because White is the reference category, the results suggest Black adults earn \$3.440 less an hour, on average, than White adults after accounting for age and gender. Hispanic adults earn, on average, \$3.766 less an hour than Whites and those of other race-ethnicities (of which Asian is the largest category) earn, on average, \$6.664 more than White adults.

Limitations of traditional approaches. A key and well-established limitation of the above approach is that we do not know how non-reference category groups compare to each other. For example, do Black and Hispanic adults differ in their hourly wages? A related issue is we do not have any real understanding of the effect of race-ethnicity as an overarching construct. Instead, we only have information on three contrasts with a single reference group (Whites).

Another recently detailed limitation is the often uncritical choice of the reference category (Johfre and Freese 2021). For example, we chose to omit White as the reference category above to illustrate what we see as the traditional approach: omit the dominant and/or or majority group as the reference. However, this approach can further reify that group as the correct or normative reference by which all others are judged. In addition, it can obscure differences among non-reference groups if analysts do not move beyond the traditional approach of relying only on the coefficients from the model to understand the nominal IV's effect. Our inequality approach of summarizing effects detailed below obviates the reference category problem entirely.

A final limitation is this approach uses the model coefficients to determine effect sizes. In nonlinear/categorical models, the coefficients summarize effects in a different metric than the "natural metric" of the dependent variable. For example, in a binary logit, the coefficients are in the metric of log odds—or odds if exponentiated—while the more natural metric of a binary outcome is the predicted probability (Mize 2019; Long and Mustillo 2021). In addition to interpretation limitations, the coefficients are problematic in comparisons across models such as in tests of mediation or in comparisons across groups (Karlson, Holm, and Breen 2012; Breen, Karlson, and Holm 2018; Long and Mustillo 2021; Williams and Jorgensen 2022). Our *ME inequality* measure utilizes predictions in the natural metric, so it does not have this limitation.

Pairwise Comparisons

One approach of interpretation that we believe has much merit but also drawbacks is to examine all possible pairwise comparisons between the categories of the nominal or ordinal IV. The benefits of this approach are there is no reference category and it provides a comprehensive understanding of how all categories compare. Drawbacks are that the number of comparisons becomes cumbersome quickly and significance testing errors become likely, and it does not provide a holistic understanding of the IV's overall effect.

For a nominal or ordinal IV with L total categories:

$$\#\ \textit{of pairwise comparisons} = \frac{L(L-1)}{2}. \tag{3}$$

For example, with the four-category race-ethnicity variable used earlier, there are $4 * 3/2 = 6$ comparisons. To calculate these comparisons, we first need to make predictions in the metric of interest. We use η (eta) throughout to refer to the prediction of interest. For this example, we are using a linear model in the metric of hourly wages so do not need to transform our predictions but can simply solve for $\mathbf{x}_i\hat{\boldsymbol{\beta}}$ using the estimates in Equation 2 to calculate predicted wages (η).

Throughout, we use x_k to refer to a focal IV and $\mathbf{x}_{-k}$ to refer to other variables in the model, such as control variables. For a focal nominal IV x_k with L total categories, we estimate comparisons among each category (denoted as a and b):

$$\textit{pairwise comparisons} = \eta\left(x_k = a, \mathbf{x}_{-k} = \mathbf{x}^*_{-k}\right) - \eta\left(x_k = b, \mathbf{x}_{-k} = \mathbf{x}^*_{-k}\right); \textit{for all } b > a. \tag{4}$$

Where control variables are held at specific values $\mathbf{x}^*_{-k}$, such as their means or at observed values (see Long and Freese 2014; Mize and Han Forthcoming). We can then use post-estimation Wald tests to determine the significance of the contrasts. Table 1 shows the predicted wages for each of the four racial-ethnic groups along with the tests of pairwise comparisons of each group. In addition to the three contrasts we already knew from the coefficients themselves, we learn that Black and Hispanic adults earn statistically equivalent wages. All other contrasts are statistically different, and some are quite large in substantive size, for example, the gaps between those of other race-ethnicities and Black adults and the gaps between those of other race-ethnicities and Hispanic adults are both over $10 an hour.

Table 1: Predicted hourly wages ($) and pairwise comparisons of predictions for racial-ethnic groups.

(1) Predicted Hourly Wages ($)		(2) Pairwise Comparisons	(3) Contrast
White	23.479	Black vs. White	$-3.440^{\dagger}$
			(1.351)
Black	20.039	Other vs. White	6.664^{*}
			(1.705)
Other	30.143	Hispanic vs. White	-3.766^{*}
			(1.294)
Hispanic	19.713	Other vs. Black	10.104^{*}
			(2.081)
		Hispanic vs. Black	-0.326
			(1.753)
		Hispanic vs. Other	-10.430^{*}
			(2.032)

Note: Standard errors in parentheses. $^{\dagger}p < 0.05$, $^{*}p < 0.01$ in a two-tailed test.

Although this approach is fairly manageable for an IV with few categories, it becomes difficult quickly. For example, one of the GSS's uncollapsed race-ethnicity variables (*racecen1*) includes 16 categories: for this variable, there would be $16^{*}15/2 = 120$ comparisons. This number of comparisons makes summarizing effects practically unreasonable and introduces statistical significance testing concerns. This is part of the motivation for our new summary measures, which result in a single test of a nominal or ordinal IV's holistic effect.

Corrections for multiple significance tests. An additional issue with calculating all pairwise comparisons—on top of the difficulty of understanding so many contrasts—is that some of the contrasts may be statistically significant by chance alone (i.e., they may be type I errors or "false positives"). It can be tempting then to use a multiple testing correction for the *p*-values to ensure that the false positive rate is 0.05 (or any desired level) across *all* tests instead of across each individual test (which will lead to much higher than a 0.05 false positive rate across all tests). For example, with the 16-category race-ethnicity variable mentioned above which produces 120 comparisons, we would have at least one false positive about 99.8 percent of the time (Curran-Everett 2000).[1]

A practical issue with such multiple testing corrections is that there is little agreement as to whether these corrections should be done and what correction is best if you do decide to implement them (Lazic 2024). Some of the most well-known corrections (e.g., the Bonferroni correction) are generally viewed poorly by methodologists and seen as overly strict. In addition, the corrections make individual significance tests on each contrast nonsensical and are often statistically motivated by a different desire than the substantive motivation of understanding the individual contrasts (Rothman 1990; Perneger 1998; Gelman, Hill, and Yajima 2012; Althouse 2016). Importantly, our proposed *ME inequality* summary measure builds on the pairwise comparison approach but avoids this unsettled issue of multiple comparisons as it always produces a single *ME inequality* estimate, regardless of the number of nominal IV categories.

Testing hypothesized contrasts only. A partial pairwise comparison approach would include testing only the hypothesized contrasts between specific categories. For example, it is extremely unlikely that a researcher would have defendable hypotheses about all 120 pairwise comparisons possible from the 16-category race-ethnicity variable in the GSS mentioned earlier. However, the researcher may have a well-motivated hypothesis about the difference between Black and American Indian individuals.[2] In this case, this contrast (and others if hypothesized) could be tested directly and all other pairwise comparisons ignored or at least not emphasized.

In general, we do not think that hypothesized contrasts need to be subjected to a multiple comparison procedure as these are individual tests of separate hypotheses. The logic of multiple hypothesis corrections is most appropriately applied when a researcher wants to know whether any one of many tests is significant. Applying a correction to determine if *a given contrast* is significant does not match the logic of the correction and will provide an incoherent answer (Perneger 1998; Althouse 2016; Rubin 2021, 2024; García-Pérez 2023). Rubin (2024) provides an accessible discussion of these issues with recommendations.

In general, we have no issue with testing specific pairwise comparisons if justified theoretically or by some other substantive motivation. Indeed, we believe that our approach for a single inequality summary measure could be combined with additional tests of specific hypothesized pairwise comparisons to good effect.

Other Existing Measures of Nominal and Ordinal IV Effects

In this section, we provide a brief overview of some of the other existing methods for summarizing effects for nominal and ordinal IVs. We begin with some alternative ways to quantify contrasts among nominal/ordinal IV categories and then cover single number summary measures. We end by pointing out some common limitations of existing approaches, which motivate the need for our new *ME inequality* summary measures.

Alternative Approaches Using Contrasts

The traditional approach for interpreting nominal and ordinal IVs is to examine coefficient contrasts between an omitted category and all other categories (see the Traditional nominal and ordinal IV interpretations in statistical models section). Multiple approaches build on this basic idea but attempt to make the contrast more general and informative.

Deviation from the (conditional) mean. Perhaps the most popular alternative is to express contrasts between individual categories of a nominal or ordinal IV and the mean—or conditional mean—across the sample. For example, we calculate a prediction for category l of the nominal IV and then compare that prediction to the mean in the sample:

$$mean\ contrast_l = \eta\left(x_k = l,\right) - \eta, \tag{5}$$

Table 2: Predicted hourly wages ($) along with mean and binary contrasts of predictions for racial-ethnic groups.

	(1) Predicted Hourly Wages ($)	(2) Mean Contrast ($\overline{Wages} = 23.127$)	(3) Binary Contrast
White	23.479	0.355	1.351
Black	20.039	$-3.086^{\dagger}$	$-3.416^{\dagger}$
Other	30.143	7.019^{*}	7.449^{*}
Hispanic	19.713	-3.411^{*}	-3.824^{*}

Note: $^{\dagger}p < 0.05$, $^{*}p < 0.01$ in a two-tailed test.

where η without any conditions will be the same as the sample mean of the outcome ($\overline{y}$).[3] We would then repeat this calculation for all L categories of the nominal/ordinal IV.

Column 2 in Table 2 shows the results of mean contrasts for each racial-ethnic category in the data. The results suggest that White individuals' earnings are equivalent to the average person in the sample, while those from other race-ethnicities earn more than the average person and Black and Hispanic adults earn less than the average.

Although we find mean contrasts easy to understand, there are limitations. First, all groups are included in the overall mean so the comparison category includes the focal group itself (Johfre and Freese 2021; Freese and Johfre 2022). This is especially suboptimal for larger groups. For example, the sample is 73 percent White so the mean contrast of η $(\textit{race-ethnicity} = \textit{White})$ versus the overall mean $(\eta = \bar{y})$ includes mostly White individuals in both statistics.

Binary contrasts. A recently suggested alternative to mean contrasts is binary contrasts, which similarly contrast the prediction for a given category to the rest of the sample, but remove the focal category from the reference prediction (Johfre and Freese 2021; Freese and Johfre 2022). Using our notation,

$$\textit{binary contrast}_l = \eta\left(x_k = l\right) - \eta\left(x_k = 1 \ldots L; \neq l\right), \tag{6}$$

where $\eta\left(x_k = 1 \ldots L; \neq l\right)$ is the prediction pooling over all other groups in the sample except the focal group l.

Column 3 in Table 2 presents the binary contrasts for our wages and race-ethnicity example. The largest shift in effect size from mean to binary contrasts is for the White group, because the reference group changed the most due to it now only including the 27 percent of the sample that is non-White (i.e., the mean of wages for Black, Hispanic, and other groups combined).

We like the binary contrast approach and if one wants a middle ground between presenting all pairwise contrasts and a single summary measure, we believe this is the most appealing choice. However, a single summary measure of a nominal or ordinal IV's holistic effect is useful in many cases, and it allows for additional extensions not available with binary contrasts. For example, it is unclear how to extend the binary contrast approach to tests of interactions and current software does not support such approaches (Freese and Johfre 2022).

Existing Summary Measures

Next, we discuss existing summary measures that attempt to quantify a nominal or ordinal IV's effect in a single value. We build on the basic idea of these approaches for our own *ME inequality* measures.

Joint tests and likelihood-ratio tests. Perhaps the most well-known strategy for providing a single summary test of a nominal IV's effect is a joint test of all the nominal IV coefficients in the model, where there are $L - 1$ coefficients representing each contrast with the reference category. This can be performed with a Wald test:

$$W = \left(\hat{\boldsymbol{\beta}}_{nominal\ IV} - \boldsymbol{\beta}_0\right)' \left(Cov\left(\hat{\boldsymbol{\beta}}_{nominal\ IV}\right)\right)^{-1} \left(\hat{\boldsymbol{\beta}}_{nominal\ IV} - \boldsymbol{\beta}_0\right), \tag{7}$$

where $\hat{\boldsymbol{\beta}}_{nominal\ IV}$ contains each of the coefficients for the nominal IV and $\boldsymbol{\beta}_0$ is the (null) hypothesized effects (that the coefficients are zero).

Continuing with our example of racial-ethnic differences in hourly wages, we calculate a joint Wald test for the three race-ethnicity coefficients in the model and find they are jointly significant ($F_{df=3} = 10.93$, $p < 0.01$). This is useful as a starting place to indicate that race-ethnicity does indeed have some effect but does not provide us with an effect size or understanding of the substantive significance of the IV.

A related test is the likelihood-ratio test of two models. For any two models that can be fit with maximum likelihood, their fit can be compared

$$\textit{Likelihood-ratio test} = 2\ln L_u - 2\ln L_c, \tag{8}$$

where $\ln L_c$ is the log likelihood of the constrained model without the nominal IV included and $\ln L_u$ is the full unconstrained model including the nominal IV. This provides an equivalent answer to that of the joint test described above, providing a single significance test as to whether the nominal IV has an overall effect or not (for this example, $\chi^2_{df=3} = 32.59,\ p <\ 0.01$).[4] However, it has the same limitations as well, with it unclear what the effect size is or its substantive significance.

Incremental R^2. A related measure is the incremental R^2, which quantifies the change in explained variation before and after accounting for the nominal IV. For example, if we fit a model of *wages* with only *age* and *gender* as predictors, the R^2 is 0.081 indicating that age and gender explain about 8.1 percent of the variation in wages. If we add *race-ethnicity* to the model, the R^2 increases to 0.098, which suggests that race-ethnicity explains an additional 1.7 percent of the variation in wages ($0.098 - 0.081 = 0.017$).

Although incremental R^2 values are easy to understand, they do not provide a tangible effect size estimate and the same R^2 value can be associated with different sized substantive effects. In addition, these measures are less meaningful in categorical outcome models (Long 1997; Long and Freese 2014).

Heise's sheaf coefficient. The approach most similar in spirit to our own, and from which we draw inspiration, is the sheaf coefficient approach of Heise (1972; also see

Whitt 1986). Heise (1972) proposed the sheaf coefficient as a summary measure of a nominal IVs effect in a structural equation modeling framework. Specifically, he proposed that all nominal IV coefficients, which represent specific contrasts with a reference group, influence a latent variable that is the holistic construct (n.b. "sheaf" refers to a "... bundle, cluster, or collection," https://www.dictionary.com/browse/sheaf). For example, our Black, Hispanic, and other coefficients from the earlier example would all predict a holistic construct of race-ethnicity. This approach allows for both a single significance test and a summary measure of the bundle of variables' effect.

Despite the sheaf coefficient's origins in structural equation modeling, we can calculate it from a multiple regression model using standardized coefficients. For a four-category nominal IV with three coefficient contrasts, as with our race-ethnicity example, the sheaf coefficient is

$$\begin{aligned}&\textit{sheaf coefficient}\\&= \sqrt{\hat{\beta}^2_{x_1,st.} + \hat{\beta}^2_{x_2,st.} + \hat{\beta}^2_{x_3,st.} + 2\left(\hat{\beta}_{x_1,st.}, \hat{\beta}_{x_2,st.} r_{x_1,x_2} + \hat{\beta}_{x_1,st.}, \hat{\beta}_{x_3,st.} r_{x_1,x_3} + \hat{\beta}_{x_2,st.}, \hat{\beta}_{x_3,st.} r_{x_2,x_3}\right)},\end{aligned} \tag{9}$$

where x-standardized coefficients are used and r_{x_a,x_b} is the correlation between the indicator variables for nominal IV categories a and b (Whitt 1986). Applying this formula to our *wages* and *race-ethnicity* model results in a sheaf coefficient of 2.263 ($p < 0.01$).

Despite our fondness for this general idea and appreciation of a single effect size measure produced, it does not correspond cleanly to the substantive significance of the effect. This is partially because the nominal IV's effect is represented as a continuous latent variable, leading to interpretations such as "for a standard deviation increase in race-ethnicity," which does not correspond to typical ways to think about nominal IV effects. Despite these limitations, we build on the principle that nominal IVs are holistic constructs with an effect equivalent to the sum of all the individual category effects for our own proposed measures.

Loglinear models. For aggregated categorical data, loglinear models provide an option for summarizing effects of nominal IVs. Loglinear models are used to analyze relationships in a contingency table of two or more binary or nominal/ordinal variables. As shown in the Joint tests and likelihood-ratio tests section, a likelihood-ratio test can also be used with a loglinear model to provide a single summary measure of the overall significance of a nominal IV—though here too it doesn't reflect the effect size directly (Agresti 2013).

Multiple summary measures of association have been proposed for loglinear models, which have been shown to be closely related (Bouchet-Valat 2022). For example, the Altham index and the intrinsic association coefficient both provide a single number summary of association even when an IV has multiple categories. In particular, the normalized intrinsic association coefficient is motivated similar to our proposed measures: to apply to any nominal IV, to not be systematically affected by the number of IV categories, and to be on an interpretable scale (Bouchet-Valat 2022).

A limitation of loglinear modeling in general is that it is only applicable to the case of categorical predictors with categorical outcomes. In addition, it is more difficult to incorporate control variables as they must too be categorical, and because loglinear models analyze a contingency table it becomes unwieldy quickly with multiple IVs. Finally, effects in loglinear models tend to be interpreted as odds ratios, and the summary measures discussed above are similarly based on these (logged) values. For our proposed measures, we use MEs, which represent absolute differences between groups in the natural metric of the outcome, rather than relative differences in a transformed metric as with an odds ratio.

Limitations of Current Approaches

Although all the approaches outlined thus far are appropriate for specific cases, they are often limited in their applications. For example, many are only appropriate for linear models and/or linear effects. In addition, we do not find the interpretation of most of these methods substantively satisfying. That is, most do not summarize effects in an intuitive metric that makes it easy to understand the substantive effect size in addition to the statistical significance of an effect.

Finally, methods that rely on coefficients for understanding effect sizes—as most of these do—present issues in many applications. For one, the coefficients often summarize effects in a different metric than that of most interest. For example, when applying the traditional approach outlined in the Traditional nominal and ordinal IV interpretations in statistical models section to a nonlinear/categorical model, such as a binary logit, the coefficients would be in the metric of log-odds or odds ratios (if exponentiated). If the metric of most interest is the predicted probabilities, a transformation of the predictions is needed. Increasingly, researchers prefer to work in the "natural metric" of the dependent variable with categorical outcomes—predicted probabilities in binary, nominal, and ordinal models—both because it provides a more interpretable effect size and because effects in this metric have better statistical properties (Mood 2010; Williams 2012; Mize 2019; Mize, Doan, and Long 2019; Long and Mustillo 2021). For example, it is inappropriate to compare logit/probit coefficients across models to determine differences in effect sizes but predicted probabilities can be compared without issue (Breen et al. 2018; Mize et al. 2019; Williams and Jorgensen 2022).

Next, we derive our *ME inequality* measures, which utilize model predictions. In doing so, our approach is applicable across linear and nonlinear/categorical models and provides an intuitive and substantively meaningful effect size metric. In addition, because it uses predictions in the natural metric of the dependent variable, it can be used to compare effects both within and across linear and categorical models.

Proposed ME Inequality Summary Measures of Nominal and Ordinal IV Effects

We propose a new *ME inequality* measure as a summary of the holistic effect of a nominal or ordinal IV. The core idea is that a nominal or ordinal IV has an effect size in concordance with the degree that it patterns the outcomes. The more disparate

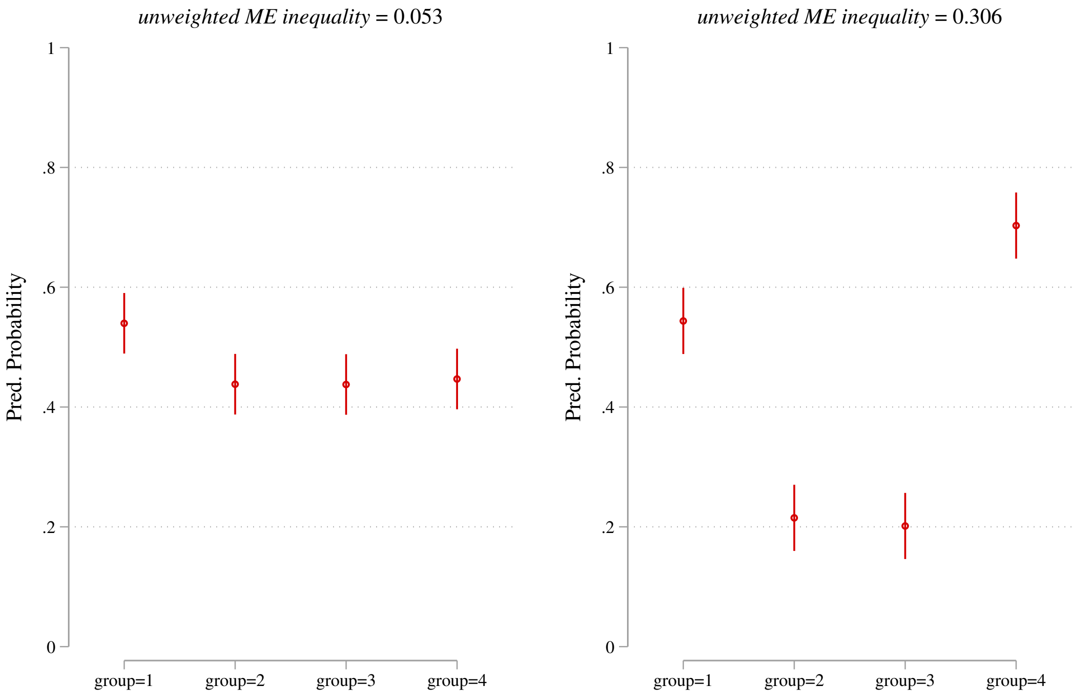

Figure 1: Hypothetical examples of low inequality (left panel) and high inequality (right panel) among categories of two different nominal IVs.

outcomes are across the categories of the nominal or ordinal IV, the larger the effect. For example, if race-ethnicity is the focal IV, then race-ethnicity's holistic effect is equal to the amount of inequality in an outcome explained by race-ethnicity. If racial-ethnic groups' outcomes are similar, there is little inequality and little effect of race-ethnicity. If racial-ethnic groups differ dramatically in their outcomes, the inequality is high and the effect of race-ethnicity as a holistic construct is large.

To illustrate the intuition behind our *ME inequality* measure, we present simulated data in Figure 1. Each panel of Figure 1 shows predictions for a different four-category IV; the outcome is binary, so the predictions are in the metric of predicted probabilities. The left panel shows an example with low inequality. Intuitively, we can see that the IV has only a small effect because the four predictions are close together. In contrast, the IV shown in the right panel has a much larger effect as the four predictions are more spread out: in other words, the IV in the right panel patterns very unequal outcomes.

The goal of our inequality measures is to provide quantifiable and testable values that summarize the visual intuition from Figure 1. We derive statistics that provides a one number summary of the effect of a nominal or ordinal IV, which provide both an easily understood effect size and the ability to compare effects within and across models.

Average ME Inequality Measure

First, a model is fit, and parameters are estimated. Any regression model or another model that produces predictions for each category of the nominal/ordinal IV is amenable to our method. For example, we could use a binary logit model:

$$\eta = \Pr(y=1|\mathbf{x}) = \frac{\exp\left(\mathbf{x}\hat{\boldsymbol{\beta}}\right)}{1+\exp\left(\mathbf{x}\hat{\boldsymbol{\beta}}\right)}. \tag{10}$$

The coefficients $\hat{\boldsymbol{\beta}}$ from Equation 10 will be in the metric of log odds. We are interested in the natural metric of the outcome, the predicted probability, so make predictions for η using Equation 10. For a focal nominal IV x_k with L total categories, we make L number of predictions: one for each category of x_k. We then calculate a pairwise comparison between each prediction, which is referred to as a marginal effect (ME)—that is, a difference in predictions. For example, for a nominal or ordinal IV with three categories:

$$\begin{aligned} ME_{nominal\,1\ vs.\ 2} &= \eta\left(x_k=2,\mathbf{x}_{-k}=\mathbf{x}^*_{-k}\right)-\eta\left(x_k=1,\mathbf{x}_{-k}=\mathbf{x}^*_{-k}\right),\\ ME_{nominal\,1\ vs.\ 3} &= \eta\left(x_k=3,\mathbf{x}_{-k}=\mathbf{x}^*_{-k}\right)-\eta\left(x_k=1,\mathbf{x}_{-k}=\mathbf{x}^*_{-k}\right),\\ ME_{nominal\,2\ vs.\ 3} &= \eta\left(x_k=3,\mathbf{x}_{-k}=\mathbf{x}^*_{-k}\right)-\eta\left(x_k=2,\mathbf{x}_{-k}=\mathbf{x}^*_{-k}\right). \end{aligned} \tag{11}$$

One additional complication is where to hold the control variables in $\mathbf{x}_{-k}$ when making the predictions. In nonlinear/categorical models, predictions will differ depending on where the control variables are held. As the default, we advise researchers to use one of the two dominant methods for calculating MEs: holding the control variables at their means or averaging over observed values. Our personal default is to average over observed values, which produces an average ME (Long and Freese 2014; Mize and Han Forthcoming).[5] Alternatively, control variables could be held at specific values of interest, such as group-specific means, though this is less common (Williams 2012; Long and Freese 2014; Long and Mustillo 2021).

Once all pairwise comparisons of predictions (i.e., MEs) have been calculated, the average ME inequality is straightforwardly their mean (using absolute values of each ME). Specifically, for a nominal or ordinal IV with L outcome categories, we calculate the absolute value of all non-redundant contrasts between categories and then calculate the average by dividing by the number of comparisons[6]

$$\begin{aligned} &\textit{unweighted ME inequality} \equiv |\textit{unweighted average ME inequality}|\\ &= \frac{\textit{Sum of all }|\textit{pairwise comparisons of MEs}|}{\#\textit{ of comparisons}}\\ &= \sum_{a=1}^{L}\sum_{b=2;b>a}^{L}\left|\eta\left(x_k=a,\mathbf{x}_{-k}=\mathbf{x}^*_{-k}\right)-\eta\left(x_k=b,\mathbf{x}_{-k}=\mathbf{x}^*_{-k}\right)\right| / \frac{L\left(L-1\right)}{2}. \end{aligned} \tag{12}$$

We use the absolute value of the contrasts to avoid specifying any one category as the reference. We refer to this quantity as the *unweighted average absolute ME inequality* or just *unweighted ME inequality* for short.[7] For example, for race-ethnicity, this would reflect how different each racial-ethnic group is from each other, on average, in the metric of interest. We explain why the term *unweighted* is needed for this version of the statistic in the next section.

Weighted average absolute ME inequality. A potential downside of the raw average ME inequality shown in Equation 12 is that it gives each category of the nominal IV equal weight in the calculation (i.e., it is an *unweighted* measure). Sometimes, this will be desirable (see discussion in the When to use weighted versus unweighted ME inequality section). However, in most cases, we think it is best to weight the contrasts to be proportional to the number of observations in each contrast (e.g., racial-ethnic groups with larger populations will be given more weight in the calculations to reflect their sample/population size). For example, if group a is Black adults and group b is Hispanic adults, and the data are perfectly representative of the 2024 U.S. population then 14 percent of the sample will be Black and 19 percent will be Hispanic so this contrast will receive a relative weight of 0.33 to reflect their sample proportion.

To calculate a weighted ME inequality measure, we first estimate contrast-specific weights, which represent the proportion of the sample included in each comparison[8]

$$a \textit{ vs. } b \textit{ contrast weight} = w_{a,b} = \frac{n_a + n_b}{N}. \tag{13}$$

We also include a correction for the fact that each group is represented in multiple contrasts so the total will sum to more than one; it will always sum to $L - 1$, therefore[9]:

$$\begin{aligned} &\textit{ME inequality} \equiv |\textit{weighted average ME inequality}| \\ &= \sum_{a=1}^{L} \sum_{b=2;b>a}^{L} \frac{w_{a,b}}{L-1} * \left|\eta\left(x_k = a, \mathbf{x}_{-k} = \mathbf{x}^*_{-k}\right) - \eta\left(x_k = b, \mathbf{x}_{-k} = \mathbf{x}^*_{-k}\right)\right|. \end{aligned} \tag{14}$$

Equation 14 produces a *weighted average absolute ME inequality*, which we refer to simply as *ME inequality* for short.[10] As we suggest this measure as the default, we drop the term weighted from our shortened moniker. We suggest researchers specifically add the term *unweighted* if using the version shown in Equation 12, as this will likely be less common. In contrast, if an author simply refers to *ME inequality*, it should be assumed to be the weighted version.

Inequality measures for binary IVs. Our focus in this article is on nominal and ordinal IVs to develop new methods for these types of variables. However, it is worth reminding readers that binary variables are a special case of nominal variables with only two categories, and our methods are equally appropriate for them. Consider the average absolute weighted inequality measure for a binary IV coded as 1 and 2:

$$\begin{aligned} &\textit{ME inequality}_{\textit{binary IV}} \equiv |\textit{average weighted ME inequality}|_{\textit{binary IV}} \\ &= \sum_{a=1}^{L} \sum_{b=2,b>a}^{L} \frac{w_{a,b}}{L-1} * \left|\eta\left(x_k = a, \mathbf{x}_{-k} = \mathbf{x}^*_{-k}\right) - \eta\left(x_k = b, \mathbf{x}_{-k} = \mathbf{x}^*_{-k}\right)\right| \\ &= \sum_{a=1}^{2} \sum_{b=2,b>a}^{2} \frac{1}{2-1} * \left|\eta\left(x_k = a, \mathbf{x}_{-k} = \mathbf{x}^*_{-k}\right) - \eta\left(x_k = b, \mathbf{x}_{-k} = \mathbf{x}^*_{-k}\right)\right| \\ &= \left|\eta\left(x_k = 1, \mathbf{x}_{-k} = \mathbf{x}^*_{-k}\right) - \eta\left(x_k = 2, \mathbf{x}_{-k} = \mathbf{x}^*_{-k}\right)\right|. \end{aligned} \tag{15}$$

In this case, because there are only two categories there is only one pairwise comparison and because the two categories represent the entire sample, $w_{a,b}$ is equal to one. Therefore, for a binary IV, the *ME inequality* is equivalent to the absolute value of the ME for that IV and there is no difference between the weighted and unweighted measure. This means our inequality approach is not strictly necessary and somewhat redundant with existing methods in this case. On the other hand, this means that conceptualizing effects for binary IVs as inequalities is equally valid. This suggests one can compare inequalities across binary, ordinal, and nominal IVs with equal validity.

Standard Errors and Significance Tests

Equations 12 and 14 provide the formulas for calculating the ME inequality measures but do not give details on calculating standard errors. It is important to remember that ME inequality is calculated based on model predictions, which themselves have uncertainty. That is, ME inequality is an estimate with an associated confidence interval. In all our examples below, we use the Delta method to estimate standard errors. The Delta method uses a first-order Taylor series approximation to calculate the variance of an estimate that is the function of random variables—in this case, the predictions and the *ME inequality* statistics are both algebraic functions of the estimated model coefficients (Greene 2012). An equally valid alternative is bootstrapping, where the model itself along with the predictions and *ME inequality* estimates are calculated many times in bootstrapped samples and the distribution of the estimates across bootstrapped samples is used for standard error and confidence interval calculation (Efron and Tibshirani 1993). We prefer the Delta method when available simply due to convenience and efficiency—bootstrapping is far more resource intensive and slower (Dowd, Greene, and Norton 2014). In practice, we have found the Delta method to work well in most cases but to struggle with convergence when the nominal or ordinal IV has many categories. In these cases, bootstrapping will be slower but has no issues with estimation even with many categories.

Once we have the standard error estimate, we can calculate 95 percent confidence intervals as[11]

$$95\%\ CIs = ME\ inequality \pm 1.96 * SE\left(ME\ inequality\right). \tag{16}$$

We can also calculate a z-statistic

$$z = \frac{ME\ inequality}{SE(ME\ inequality)}, \tag{17}$$

which we can then use to calculate p-values for the inequality measures.

When to Use Weighted Versus Unweighted ME Inequality

Above, we suggested using the weighted inequality (*ME inequality*) measure as the default. First, we suggest this because it is similar to how most other common statistics are calculated. For example, a simple sample mean will reflect the size of the groups in the sample (e.g., the mean wage in a representative U.S. sample will most strongly reflect the wages of White individuals). For inequality as a concept, weighting to represent group sizes reflects reality in the population: it represents how different individuals from different groups are, on average, in the population. Part of this average difference is due to how many people are in each group and thus how often they would be part of the comparison.

However, this does not mean we think the *unweighted ME inequality* should never be used. In cases where the analyst wants to simply quantify differences between groups, and the differing group sizes are irrelevant or a nuisance, the unweighted measure does not reflect group size. Similarly, sometimes the analyst will want to effectively control for group size and the unweighted measure does so. For example, if we want to compare how different religious groups were in 1972 and how different they were in 2021 as we do in the Comparing ME inequality effect sizes for a focal IV across models section, it is important to recognize that the relative size of the individual religious groups has changed. For example, the GSS finds that 64 percent of U.S. adults identified as Protestant in 1972, whereas only 40 percent do so today. Comparing weighted *ME inequality* statistics across the two time periods would reflect both the differences in outcomes among the religious groups and also shifts in the size of the groups. This reflects reality and the inequality at the population level. However, in some cases, one may want to simply quantify pure differences between groups regardless of their respective sizes, and the *unweighted ME inequality* is the appropriate statistic in this case.

Comparison to Other Measures of Inequality

Inequality as a concept is central to many social and biological sciences (Allison 1978; Schleuter et al. 2010). Existing inequality statistics are typically focused on summarizing inequality at the level of a distribution. For example, inequality statistics are often calculated to summarize income distributions. Perhaps the most popular, the Gini coefficient can be calculated as (Dagum 1998; Liao 2022)

$$Gini = \sum_{i=1}^{N} \sum_{j=2;j>i}^{N} \left| x_i - x_j \right| / 2N^2 \bar{x}, \tag{18}$$

where i and j denote individual observations in the data. Note that the numerator of Equation 18 is almost identical to that of our *ME inequality* measures shown above, but the Gini coefficient summarizes inequality at the distributional level by comparing all individual observations in the sample—while our measure compares predictions for groups. The two measures also differ in the denominator, with Gini scaled to range from 0 to 1 so that 0 represents perfect inequality and 1 represents maximal inequality.

Other conceptually similar measures come from the biological sciences in the form of diversity indices (Schleuter et al. 2010). These measures are typically designed to measure how evenly distributed groups are across the sample/population. These measures are also sometimes applied in demography to measure segregation, for example, Shannon and Simpson's index/entropy (White 1986). For example, where p_l is the proportion of the sample in category l, Simpson's index is

$$Simpson's\ index = \sum_{l=1}^{L} p_l^2, \tag{19}$$

which provides the probability that two observations selected at random will be from the same group. Similar to the Gini coefficient, these measures focus on summarizing variation at the distributional level but here are specifically designed to compare across the distribution of nominal groupings.

Although our measures of inequality take inspiration from the goals of these inequality and diversity statistics, we apply the idea at a different level of analysis. Instead of focusing on inequality in a distribution, we focus on inequality of average outcomes across nominal or ordinal groupings. In addition, our approach is quite general in that it can be applied to any model in which predictions can be made for the nominal/ordinal groupings. This allows inequality to be calculated as raw differences among groups or based on a conditional relationship, usually from a regression model where the focal nominal/ordinal IV is one of multiple predictors. In the Examples of using ME inequality summary measures section, we show a host of different model-based applications of our inequality measures.

Defining the Estimand

In this section, we have derived the *ME inequality* statistics and compared them to related measures. At this point, we can more precisely state what *ME inequality* measures substantively and how it can be interpreted. We see three equally valid interpretations. First, ME inequality can be described as the *average difference in outcomes between observations in distinct groups*. For example, in our race-ethnicity and wages example, *ME inequality* quantifies how divergent individuals of different race-ethnicities' wages are, on average. We suspect this will be the most widely applied interpretation as it implies no causality but instead reflects a simple description of average adjusted or unadjusted differences across groups.

Alternatively, *ME inequality* could be used to quantify a causal effect of a nominal or ordinal IV. In this case, *ME inequality* summarizes *how outcomes would typically change if an observation switched from their current category to another category*. For example, consider a randomized medical trial where there are four different treatments for an illness. Imagine that a participant is given treatment *A* but that it results in unpleasant side effects. *ME inequality* could quantify how we expect the participant's illness outcome to change if they were to switch to another of the three treatments. Note that the calculation of *ME inequality* is identical in each case and thus the ability to infer causality is not a product of our measure but instead of the data, model, context, and assumptions the analyst is willing to make.

Finally, our *ME inequality* measure can also be interpreted similar to a diversity index, as described in the prior subsection. In this case, it quantifies *how different two observations from different groups, picked at random, will be on average*. For example, returning to the race-ethnicity and wages example, it quantifies how different two individuals' wages would be if we randomly compared two people of different race-ethnicities.

Next, we turn to a series of applications of the *ME inequality* measures.

Examples of Using ME Inequality Summary Measures

We illustrate how to use our *ME inequality* measures in several different situations below. These cover single and multiple models and one or multiple focal IVs.

Inequality as a Summary of a Nominal or Ordinal IV's Effect

First, we illustrate the simplest case where one wants only to summarize the effect of a single nominal or ordinal IV from a single model. The *ME inequality* measure is still informative in this case if a single measure of effect size is desired.

For this example, we revisit our model from the Nominal and ordinal IV section using 2021 GSS data on *race-ethnicity* and *hourly wages*. Table 1 provides the predictions for each group along with the associated pairwise comparisons from a linear model regressing wages on race-ethnicity. Using Equation 12, we calculate the *unweighted ME inequality*:

$$\begin{aligned} &\textit{unweighted ME inequality} = \frac{\textit{Sum of all } |\textit{pairwise comparisons of MEs}|}{\textit{\# of comparisons}} \\ &= \frac{|\eta_{Black} - \eta_{White}| + |\eta_{other} - \eta_{White}| + |\eta_{Hispanic} - \eta_{White}| + |\eta_{other} - \eta_{Black}| + |\eta_{Hispanic} - \eta_{Black}| + |\eta_{Hispanic} - \eta_{other}|}{\frac{4\,(4-1)}{2}} \\ &= \frac{|-3.440| + |6.664| + |-3.766| + |10.104| + |-0.326| + |-10.430|}{6} = \frac{34.730}{6} = 5.788. \end{aligned} \tag{20}$$

For this first example, we show all calculations in Equation 20 to be clear how the statistic is calculated. The numerator sums all the pairwise comparisons among the IV categories (using their absolute value). The denominator is the total number of pairwise comparisons, which makes the final *unweighted ME inequality* statistic a simple average.

We calculate the average absolute *unweighted ME inequality* to be 5.788, which indicates that individuals from different racial-ethnic groups' wages differ by \$5.79 on average. That is, it quantifies the degree to which race-ethnicity, as a holistic construct, patterns outcomes on average.

Uncertainty of the estimate. How certain are we in the \$5.79 inequality estimate, and is this estimate significantly different than zero? Using the Delta method (see the Standard errors and significance tests section), we estimate the standard error of the *unweighted ME inequality* to be 1.042 suggesting a fairly precise estimate. Indeed, this translates to a 95 percent confidence interval of $3.747 - 7.830$ and a p-value of <0.01 for a test against a null of zero or no effect (for advice on interpreting p-values, see Wasserstein and Lazar 2016).

Effect sizes. Is \$5.79 a small or large amount of inequality? In this case, wages provide an intuitive metric and a difference of almost six dollars an hour is quite large and meaningful. An option for contextualizing effect sizes in linear models, which is especially useful when the outcome has no clear metric, is to compare the size of the inequality measure to the standard deviation of the outcome.[12] In this case, the standard deviation of wages is 17.122 and our *unweighted ME inequality* represents about 0.34 standard deviations ($5.778/17.122 = 0.337$). It is important to note that even small effect sizes can be meaningful in certain applications and the importance of any given effect size is context dependent.

Weighted inequality as a summary measure of an IV's effect. The *unweighted ME inequality* measure reported above effectively weights each racial-ethnic group equally in the calculation. Usually, it is best to give more weight in the calculations to groups that are larger, as we do in Equation 14 for our weighted *ME inequality* measure (but see our discussion in the When to use weighted versus unweighted ME inequality section). Doing so results in a weighted *ME inequality* of 4.926, somewhat smaller than the *unweighted ME inequality* of 5.788. The weighted *ME inequality* is smaller because some of the largest contrasts between groups are for comparisons of smaller groups. For example, the largest gap is between Hispanics and individuals from other racial-ethnic groups, which is relatively down weighted because these groups collectively represent only 17 percent of the sample. In contrast, the gaps between Whites and each group are relatively smaller and these are given more weight in the calculations as Whites represent 73 percent of the sample. For example, the White versus Black gap is relatively small, and these two groups collectively represent 83 percent of the sample so this contrast is given a large relative weight in the calculation of *ME inequality*.

Comparing Inequalities within a Single Model

Next, we illustrate comparing multiple inequality measures within a single model.

Comparing effects of a nominal or ordinal IV across groups. This example focuses on comparing the effects of a single nominal or ordinal IV across multiple groups in the data. We accomplish this with an interaction term between the focal nominal IV and the grouping variable.

We use data on status stereotypes of gender and sexuality groups from Mize and Manago (2018a); in this case, sexuality is the focal nominal IV and gender is a binary grouping variable we wish to compare effects across. Mize and Manago (2018b) argue that there are "…highly discrepant status distinctions among men's sexual orientation categories" but that "…women's sexual orientation groups have relatively less status differentiation" (p. 306). They base their argument on data from a survey experiment on stereotypes of the status of heterosexual men, bisexual men, gay men, heterosexual women, bisexual women, and lesbian women. We recreate the figure showing status stereotypes from Mize and Manago (2018b) here as Figure 2.

The left side of Figure 2 shows large differences among the sexuality groups for men, driven by the very high rating of heterosexual men's status. In contrast, among the sexuality groups for women on the right side of Figure 2, there appears to be relatively less status differentiation.

We calculate *ME inequality* for the sexuality groupings separately for men and women.[13] We find that among men, the sexuality groups differ by 1.173 points on the status scale on average, which is almost one standard deviation. Among women, the *ME inequality* is 0.515, notably smaller at less than half the size of the

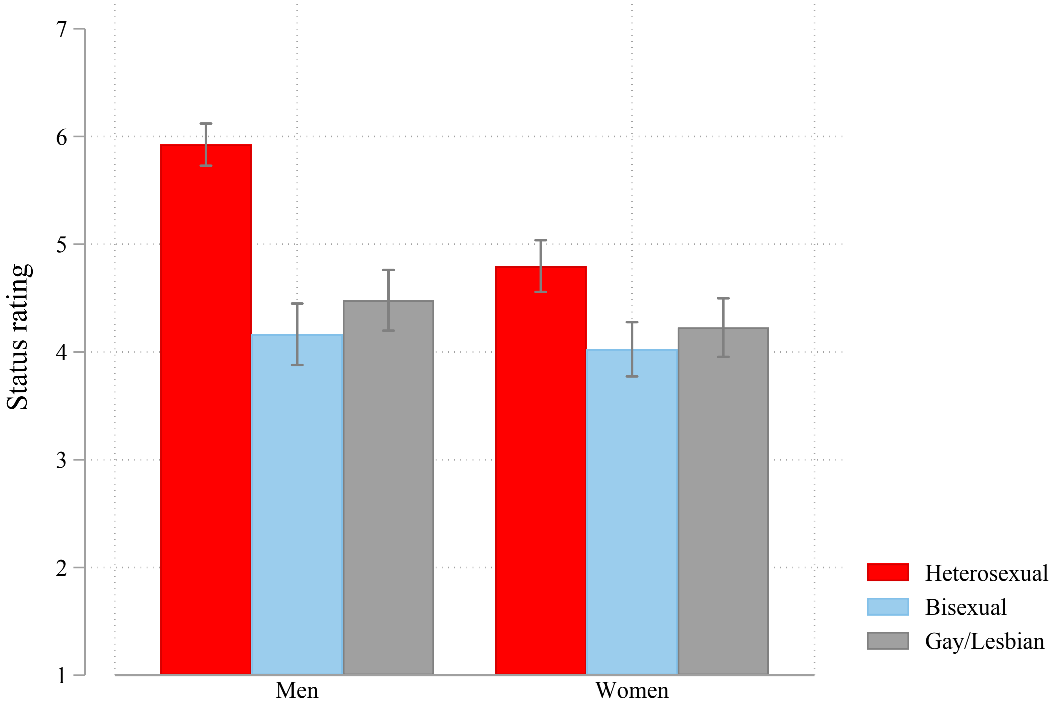

Figure 2: Perceived status of sexuality groups, separately for men and women (data from Mize and Manago 2018b).

inequality for men. We then use a Wald test to compare the two inequalities and find that the *ME inequality* for sexuality groups is larger for men than it is for women ($\mathit{ME\ inequality}_{men} - \mathit{ME\ inequality}_{women} = 1.173 - 0.515 \approx 0.659, p < 0.01$), which supports Mize and Manago's (2018b) original claims but with the addition of a testable statistical hypothesis and a quantified effect size.

Comparing across different binary, nominal, and ordinal IVs. Next, we compare effects across multiple focal IVs within a single model. We use 2021 GSS data to determine the relative influence of *gender*, *race-ethnicity*, and *class* on political views. For simplicity, we use a binary outcome measure of whether the participant self identifies as *conservative* or not and fit a binary logit.

Table 3 presents the predicted probabilities from the model along with the *ME inequality* statistics. *Gender* is measured as a binary in the data, so the *ME inequality* measure is equal to the absolute value of the ME of gender, which is 0.070, indicating that men are about 7 percentage points more likely to identify as conservative than are women. *Race-ethnicity* includes four categories with an $\mathit{ME\ inequality}_{race\text{-}ethnicity}$ of 0.108, indicating that individuals from different racial-ethnic groups differ by about 11 percentage points on average in terms of their political views. Finally, *class* is an ordinal measure with four categories but because this ordinal variable is entered into the regression models as nominal, the $\mathit{ME\ inequality}_{class}$ is calculated in the same manner as for a nominal variable and equals 0.012, which is not significantly different than zero ($p = 0.489$).

Table 3: Predicted probability of identifying as conservative and *ME inequality* statistics for gender, race-ethnicity, and class.

	(1) Pr(Conservative)	(2) *ME Inequality*
Gender		
Man	0.354	0.070^*
Woman	0.285	(0.016)
Race-ethnicity		
White	0.350	0.108^*
Black	0.189	(0.013)
Hispanic	0.262	
Other	0.238	
Class		
Lower class	0.323	0.012
Working class	0.312	(0.017)
Middle class	0.319	
Upper class	0.296	

Note: Standard errors in parentheses. $^{\dagger}p < 0.05$, $^{*}p < 0.01$ in a two-tailed test.

Taken together, our results suggest that race-ethnicity and gender have meaningfully large effects on political views. In contrast, class has almost no unique effect on political views. A benefit of this approach is the ability to compare effect sizes across binary, nominal, and ordinal IVs and to compare regardless of the number of categories of the IV. In each case, a single *ME inequality* per IV can be calculated and then compared in a comparable metric. For example, a Wald test can be used to test the equality of the ME inequality statistics. In this example, we find that $\textit{ME inequality}_{\textit{race-ethnicity}}$ is significantly larger than $\textit{ME inequality}_{\textit{class}}$ (difference $= 0.096$, $p < 0.01$).

Comparing effect sizes across nominal/ordinal and continuous IVs. The prior example compares effect sizes across categorical IVs. In some situations, it is desirable to compare the effect of a categorical IV to that of a continuous IV, which presents challenges due to the differing metrics. Gelman (2008) proposes a method of comparing effects of binary IVs to continuous IVs, which Mize and Manago (2022) adapted for MEs and we build on here. Specifically, Gelman (2008) suggests that the effect of a two standard deviation increase in a continuous IV should generally be comparable to that of a binary IV. The logic is that for a binary variable with mean (μ) 0.5, the standard deviation will also be 0.5 ($SD_{binary} = \sqrt{\mu\,(1-\mu)}$) and thus two standard deviations reflect the entire range of the binary variable (from 0 to 1). For our purposes, the *ME inequality* statistic can be used as the comparable effect size measure for a nominal or ordinal IV and a +2 SD ME can be calculated to quantify effects for continuous IVs.

We build on our example from the last section predicting conservative self-identification by adding *age* and age^2 to the model.[14] Figure 3 shows predictions across the range of age observed in the data. The mean age in the data is 52.26 and the SD is 17.23. To calculate a $+2$ SD ME, we center the change on the mean

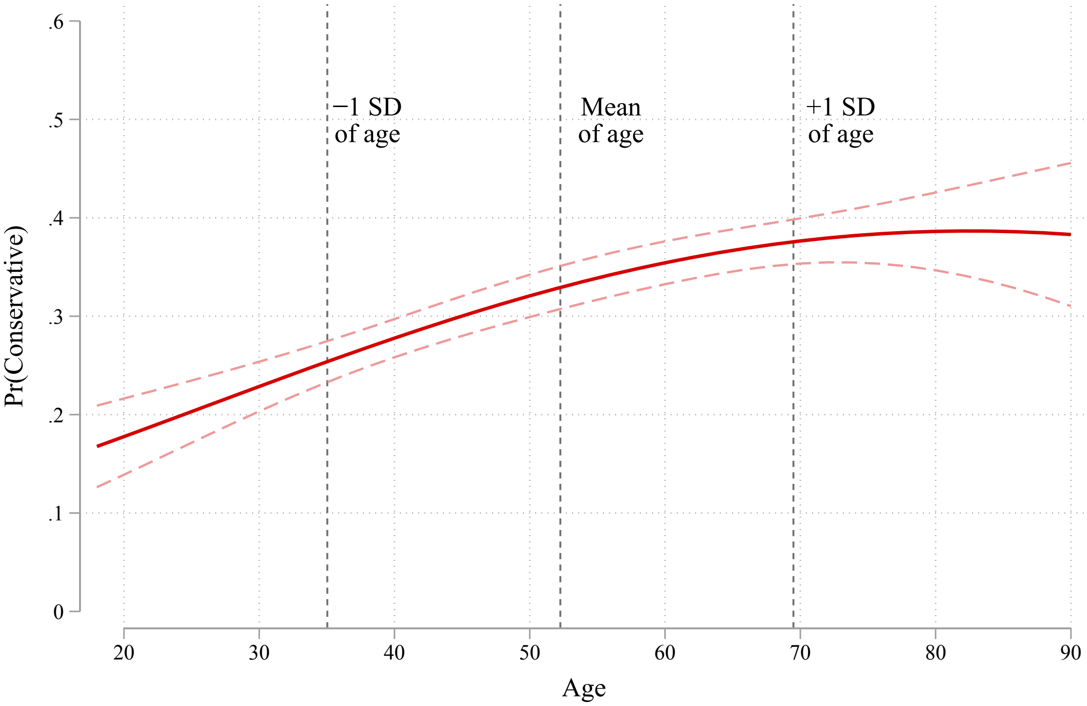

Figure 3: Predicted probability of self-identifying as conservative by age.

and calculate the effect of aging from one SD below the mean ($52.26 - 17.23 = 35.03$) to one SD above the mean ($52.26 + 17.23 = 69.49$); vertical lines in Figure 3 illustrate where the predictions are calculated. Note this is a centered change, which we recommend for such calculations as it better represents the distribution of the continuous IV than an uncentered change (Long and Freese 2014). We hold the control variables at their means $\bar{\mathbf{x}}_{-k}$ for the calculation but we could alternatively average over observed values (see Long and Freese 2014; and Mize and Han Forthcoming, for details on calculating MEs for continuous IVs)

$$\begin{aligned}&ME_{age+2SD} = \eta\left(age = \overline{age} + SD_{age}, x_{-k} = \bar{\mathbf{x}}_{-k}\right) - \eta\left(age = \overline{age} - SD_{age}, \mathbf{x}_{-k} = \bar{\mathbf{x}}_{-k}\right)\\&= \eta\left(age = 69.49, \mathbf{x}_{-k} = \bar{\mathbf{x}}_{-k}\right) - \eta\left(age = 35.03, \mathbf{x}_{-k} = \bar{\mathbf{x}}_{-k}\right)\\&= .372 - .249 = .123.\end{aligned} \tag{21}$$

The effect of a two standard deviation increase in age is to increase the probability of identifying as conservative by 12.3 percentage points ($ME_{age+2SD} = 0.123$). This effect is similar in size though slightly larger than the *ME inequality* for race-ethnicity, which is now 0.095 in the current model with age included as a predictor. A Wald test suggests no significant difference in the size of these two effects (difference $p = 0.229$).

While we generally like Gelman's (2008) method for comparing effects, we find two standard deviations a principled yet still somewhat arbitrary amount for summarizing effects of continuous IVs. For example, the logic of two standard deviations reflecting the entire range of a binary IV only holds when the mean of the binary IV is 0.5 and can represent a much smaller range for skewed binary IVs. For example, for a binary IV with a mean of 0.95, the SD is 0.218 and thus two SDs is only 0.436, or less than half of the range of the binary IV, and a +4.587 SD increase would be needed to represent the entire range—not +2 SDs.

An alternative if the goal is to always represent the range of all IVs is to calculate MEs across the range of a continuous IV. The logic is: the effect across the range of a binary variable is an increase from 0 to 1 (that is its entire range regardless of its SD). Therefore, a comparable effect for a continuous IV is an effect across its entire range (e.g., an effect of aging from the youngest to the oldest age in the data). In practice, it is useful to account for outliers and to make predictions where there is more support of the data by using a trimmed range of the 5th to 95th percentile or the 10th to 90th percentile.

To illustrate, we calculate an ME across the trimmed range of age from the 5th percentile (25) to the 95th percentile (79) in the data

$$
\begin{aligned}
&ME_{\textit{age, 5th to 95th percentile}} = \eta\left(age = 79, \mathbf{x}_{-\mathrm{k}} = \bar{\mathbf{x}}_{-\mathrm{k}}\right) - \eta\left(age = 25, \mathbf{x}_{-\mathrm{k}} = \bar{\mathbf{x}}_{-\mathrm{k}}\right) \\
&= .383 - .198 = .184.
\end{aligned}
\tag{22}
$$

The effect across the trimmed range of age is to increase the probability of identifying as conservative by 18.4 percentage points, a very large effect and one notably larger than the effect of race-ethnicity, class, or gender. For example, compared to the second largest effect, which is for race-ethnicity ($\textit{ME inequality}_{\textit{race-ethnicity}} = 0.095$), the effect of age across its trimmed range is roughly twice as large (0.184 vs. 0.095, $p < 0.01$).

Comparing Inequalities across Models

In this section, we illustrate comparisons of our *ME inequality* measures across models. In all cases, multiple models are fit, inequality statistics are calculated, and effects are then compared. One step that is necessary but that we do not cover here is combining the model estimates to allow for statistical comparisons of effects. We are using the seemingly unrelated estimation method (Wessie 1999) of comparing MEs developed by Mize et al. (2019) to combine the estimates from the models to allow for tests comparing the sizes of multiple *ME inequality* statistics. For an accessible introduction to these methods, we recommend Williams and Jorgensen (2022); full details can be found in Mize et al. (2019).

Comparing ME inequality effect sizes for a focal IV across models. First, we illustrate comparing effects of a single focal nominal IV across different models. Consider the question "has the role of religion in patterning social attitudes decreased over time?" We use an example on attitudes about *LGB free speech rights*, comparing the effect of *religious affiliation* in the 1970s to the effect of *religious affiliation* in more recent years (2010 to 2021). We fit two separate binary logits using multiple years of GSS data. Specifically, we fit the first model using data from the 1970s and the second model using data from 2010 to 2021; we also control for *age* and *gender* in each model. We then calculate the *unweighted ME inequality* for religious affiliation in both models and compare the effects.

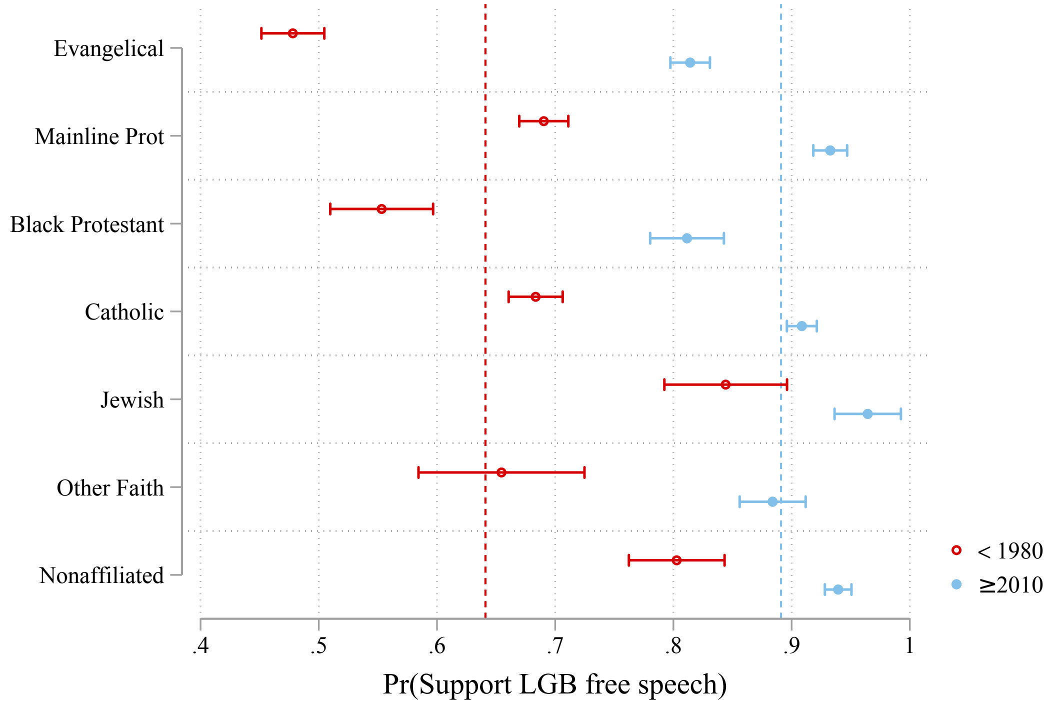

Figure 4: Predicted probability of supporting free speech rights for LGB people based on religious affiliation, in a sample from the 1970s and a sample from 2010 to 2021.

Figure 4 shows the predicted probability of supporting LGB free speech rights; religious groups are represented as separate predictions in separate rows in the plot. The predictions from the separate models for different years of data are represented with separate colors with the earlier years in red and the more recent years in light blue; vertical dashed lines show the mean in each time period.

Visually, it appears that there is more inequality in the pre-1980 period (i.e., the predictions are more spread out across groups). We estimate the *unweighted ME inequality* for the <1980 period to be 0.156 and the *unweighted ME inequality* for the ≥2010 period to be 0.072. The difference between the two time periods is 0.083 ($p < 0.01$), indicating the effect of religion has declined from the pre-1980 period to the more recent period (2010 to 2021).

Note this example represents an interaction effect (*religious affiliation * survey year*) and could be tested either using interaction terms in a single model or by fitting separate models for each time period, as we have done here. With separate models, the differences in effects can reflect both differences in the effect of the focal IV and also differences in effects of control variables, which are allowed to vary in separate models (see Long and Mustillo 2021; Blackwell and Olson 2022; Mize and Han Forthcoming). A related issue is whether the distribution of the focal nominal/ordinal IV is allowed to vary or not across models. With the *unweighted ME inequality* as shown above, the effect quantifies raw differences between religious affiliations and does not incorporate the sizes of the groups or their relative composition, which have changed over time. Alternatively, the weighted *ME* *inequality* would incorporate information on the distribution of religious affiliations in each sample, and thus the statistic will both reflect the size of the groups—to give more weight to larger groups—and also to reflect changing religious composition in the population as the weights are sample/model dependent (see the When to use weighted versus unweighted ME inequality section).

Tests of mediation/attenuation with a focal nominal or ordinal IV. Another application for using an inequality summary measure is for tests of mediation/attenuation when the focal IV is nominal or ordinal. For example, we could ask "How much of the racial-ethnic inequality in health is explained by socioeconomic factors?" Here, we could calculate *ME inequality* in a base model without the mediators (SES factors) and then also calculate *ME inequality* in a model with the mediators included and compare them using the same methods as in the last section (Mize et al. 2019). This is akin to the traditional "difference in coefficients" approach to examining mediation as effect attenuation across models, though we note that our inequality approach could also be used in the context of a causal mediation analysis (MacKinnon et al. 2002; Nguyen, Schmid, and Stuart 2021).

For this example, we use Health and Retirement Study data from 2012, which is a sample of older adults in the United States. For our outcome, we use a count of physical limitations ranging from 0 (no limitations) to 10 (limited across all ten activities assessed). Our focal nominal IV is a four-category *race-ethnicity* variable. Our mediating variables are SES factors, specifically whether the respondent has a *college* degree, their *household income*, and their *total assets*. We calculate the *ME inequality* for *race-ethnicity* in models with and without the SES factors and then compare the effects.

As our outcome is a count variable, we use a negative binomial regression model.[15] Our prediction of interest is the expected count or rate, which is calculated as

$$E\left(y|\mathbf{x}_i\right) = e^{\mathbf{x}_i\hat{\boldsymbol{\beta}}}. \tag{23}$$

For our case, the prediction for each group represents the expected number of physical limitations for someone of that race-ethnicity. Figure 5 presents the results. Predictions from the base model with no control variables are shown in red. Predictions from the model that includes the mediating SES factors are shown in light blue. Visually, it appears that the inequality among racial-ethnic groups is reduced after accounting for SES, with the blue bars on average closer together than the red bars.

To test whether SES mediates the effect of race-ethnicity, we calculate and compare *ME inequality* measures in each model. In the base model, the *ME inequality* for race-ethnicity is 0.383, indicating that racial-ethnic groups differ by about 0.4 physical limitations on average. After accounting for the SES factors, the *ME inequality* for race-ethnicity reduces to 0.217. The effect of race-ethnicity is significantly smaller after accounting for SES factors (difference $= 0.383 - 0.217 = 0.166, p < 0.01$), suggesting that SES partially mediates the effect of race-ethnicity on health; SES explains about 43 percent of the inequality in outcomes due to race-ethnicity.[16]

There are multiple benefits of the approach to mediation/attenuation shown in this section. First, it involves only one statistical test instead of countless tests for

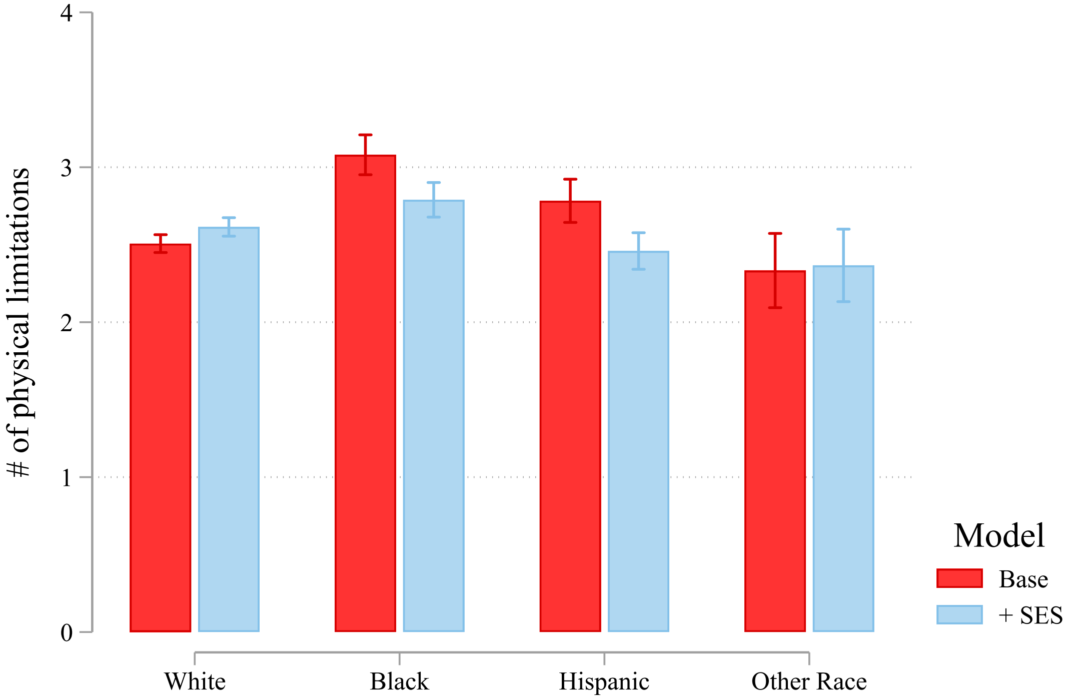

Figure 5: Predicted number of physical health limitations for different racial-ethnic groups. Base model includes no control variables; second model adds SES factors of education, income, and assets.

each contrast between racial-ethnic categories making it (a) easier to understand and present and (b) obviating the need for a correction for multiple hypothesis tests. Second, multiple mediating variables are easily incorporated—as demonstrated in this example—unlike multiple classic approaches that allow for only a single mediating variable. Third, mediating variables of any type can be used, unlike many approaches, which only allow continuous or binary mediating variables. Lastly, because *ME inequality* is calculated from the MEs for the IV, it does not suffer from issues of model rescaling that affect the coefficients (Karlson et al. 2012; Williams and Jorgensen 2022).

Other Potential Applications of ME Inequality Measures for IVs

There are several other potential applications of our *ME inequality* measures. First, the methods shown in this section can be applied to any model that results in a single prediction of substantive interest. We have shown linear, binary, and count outcome models as examples, but many other models and/or predictions also qualify, and inequality could be calculated in the same way as shown in this section. In the next section, we detail special considerations for multiple outcome models, such as those for nominal and ordinal dependent variables.

All our examples in this section use single-level regression models. However, *ME inequality* can also be calculated for multilevel and longitudinal models. In these cases though, some special considerations are needed for making accurate predictions on which the inequality statistic will be based (for details and advice, see Bland and Cook 2019; Mize and Han Forthcoming).

Another application could be to quantify the total effect of multiple IVs combined. For example, if *educational degree*, *income quartile*, and *occupational sector* are all IVs in a model a researcher may want to quantify the total effect of socio-economic status (SES), which reflects all three variables' effects. In this case, the single combined *ME inequality* measure to summarize the effect of SES is straightforwardly the sum of the *ME inequalities* for each of the three constituent IVs.[17]

In the Comparing ME inequality effect sizes for a focal IV across models section, we showed an example of comparing inequalities across different time periods. An identical empirical strategy could be used to compare inequalities across groups. In either case, the weighted *ME inequality* measure and the *unweighted ME inequality* measure test slightly different questions. A test of the equality of the two different inequality measures could help determine the reasons for similarities or differences in inequality across the samples/groups. For example, if the weighted *ME inequality* differs but the *unweighted ME inequality* does not, this suggests that the changes in inequality are due to distributional changes for the nominal/ordinal IV composition across samples/groups. This is similar in spirit to a decomposition analysis (Fortin, Lemieux, and Firpo 2011).

One additional potential application for measures of *ME inequality* is as justification for simplifying a nominal/ordinal IV into a specification with fewer categories, collapsing similar responses into one category. For example, collapsing "strongly agree" and "agree" into one category. In this case, a model could be fit with the nominal IV including all categories and a second model fit with the simplified version. If the inequality measure is similar—and/or statistically indistinguishable—between the two specifications, this could help justify the simpler measure. One note of caution, however, is that just because the inequality measure—a holistic effect quantification—is similar does not mean that all pairwise comparisons will be similar across specifications.

Nominal and Ordinal Dependent Variables

We have focused on nominal and ordinal IVs to this point, but nominal and ordinal dependent variables pose similar issues, with many effects produced often resulting in a need for a summary measure. In this section, we extend our ideas to summarizing effects for nominal and ordinal outcomes. We first cover continuous and binary IVs in these models and develop a *total ME* summary measure. Then, we cover the case of nominal/ordinal outcomes with nominal/ordinal IVs, which presents some special considerations.

Predictions in Nominal and Ordinal Outcome Models

With a nominal or ordinal outcome model—for example, the multinomial logit model or the ordered logit/probit—there are as many predicted probabilities as there are outcome categories. That is, for a nominal/ordinal outcome with Λ

categories, there are Λ predicted probabilities for both a nominal and ordinal model.

For a multinomial logit model (for a nominal DV), we can calculate Λ predicted probabilities for each outcome category λ:

$$\eta_\lambda = \Pr(y = \lambda) = \frac{\exp\left(\mathbf{x}_i \hat{\boldsymbol{\beta}}_{\lambda|\Lambda}\right)}{\sum_{\lambda=1}^{\Lambda} \exp\left(\mathbf{x}_i \hat{\boldsymbol{\beta}}_{\lambda|\Lambda}\right)}. \tag{24}$$

For an ordinal logit model, we can similarly calculate Λ predicted probabilities for each outcome category λ:

$$\eta_\lambda = \Pr(y = \lambda) = \frac{\exp\left(\mathbf{x}_i \hat{\boldsymbol{\beta}} - \tau_{\lambda-1}\right)}{1 + \exp\left(\mathbf{x}_i \hat{\boldsymbol{\beta}} - \tau_{\lambda-1}\right)} - \frac{\exp\left(\mathbf{x}_i \hat{\boldsymbol{\beta}} - \tau_{\lambda}\right)}{1 + \exp\left(\mathbf{x}_i \hat{\boldsymbol{\beta}} - \tau_{\lambda}\right)}. \tag{25}$$

In both cases, for nominal and ordinal outcome models, we have as many predictions as we have outcome categories (Λ). In these models, the metric for the prediction of interest (η) is usually the predicted probability of a given outcome category (λ), as shown in the two previous equations (denoted as η_λ). Therefore, in this case where predicted probabilities are of interest, for the purposes of calculating MEs from the predictions and summary measures based on the MEs, the models present identical challenges.

MEs and Total ME Summary Measures in Nominal and Ordinal Outcome Models

For a continuous or binary IV in nominal and ordinal outcome models, there are as many MEs as there are outcome categories (Λ). A general formula for an ME for these models is

$$ME_\lambda = \eta_\lambda\left(x_k = end, \mathbf{x}_{-k} = \mathbf{x}^*_{-k}\right) - \eta_\lambda\left(x_k = start, \mathbf{x}_{-k} = \mathbf{x}^*_{-k}\right), \tag{26}$$

where subscript λ indicates that there is a separate ME for each outcome category (ME_λ). For a focal continuous IV x_k, we pick a starting value of interest, for example, the mean ($\bar{x}_k$), and then calculate a change of Δ units

$$ME_{\lambda,+\Delta} = \eta_\lambda\left(x_k = \bar{x}_k + \Delta, \mathbf{x}_{-k} = \mathbf{x}^*_{-k}\right) - \eta_\lambda\left(x_k = \bar{x}_k, \mathbf{x}_{-k} = \mathbf{x}^*_{-k}\right), \tag{27}$$

where Δ can be any unit of change but is commonly one or a standard deviation. For a binary IV, there is no choice of which two values to calculate the ME: predictions are made at the observed values of the variable (usually zero and one) and compared

$$ME_{\lambda,binary\ IV} = \eta_\lambda\left(x_k = 1, \mathbf{x}_{-k} = \mathbf{x}^*_{-k}\right) - \eta_\lambda\left(x_k = 0, \mathbf{x}_{-k} = \mathbf{x}^*_{-k}\right). \tag{28}$$

Although calculation of the MEs is straightforward even in nominal and ordinal outcome models, the often large number of MEs produced can be cumbersome and a summary measure is useful.

A total effect summary measure. With a linear or binary outcome, a single ME summarizes the total effect of a continuous or binary IV. Similar logic can be applied in nominal and ordinal outcome models by aggregating the effect of a single IV across each outcome category. That is, an IV's *total effect* is its holistic effect across all outcomes. Then, the *total ME* statistic is simply the sum of all the MEs for that IV (using absolute values) divided by two

$$total\ ME = \sum_{\lambda=1}^{\Lambda} \left| \eta_\lambda \left(x_k = end, \mathbf{x}_{-k} = \mathbf{x}^*_{-k} \right) - \eta_\lambda \left(x_k = start, \mathbf{x}_{-k} = \mathbf{x}^*_{-k} \right) \right| /2. \quad (29)$$

The *total ME* is divided by two to account for the fact that MEs in multi-outcome models are symmetric/zero sum. That is, an increase in one outcome necessarily suggests a decrease in other outcomes (and therefore the *total ME* without taking absolute values will always equal zero). Therefore, the raw *total ME* when using absolute values of the MEs has a maximum possible value of 2. By dividing it by two, the range of the statistic is restricted to 0 to 1.

The *total ME* represents the proportion of observations that are predicted to change outcomes given a change in the IV. Put another way, this represents the complete effect of the IV across all outcome categories. The max possible effect is one, meaning a change in the IV predicts all observations to shift outcome categories.

Although the *total ME* can be calculated for specific nominal IV pairwise comparisons, it is primarily intended for use with continuous and binary IVs. For example, the pairwise comparisons of predictions for categories B versus D of a nominal IV could be accumulated using Equation 29 for a $total\ ME_{nominal\ IV,\ B\ vs.\ D}$. However, this would result in many *total MEs* and have the same issues with pairwise comparisons and multiple hypothesis tests as discussed in the Pairwise comparisons section. In the MEs for nominal or ordinal IVs in nominal and ordinal outcome models section, we discuss an approach that combines the *total ME* idea with our *ME inequality* measures discussed earlier, which we believe is more useful in most cases with both a nominal/ordinal IV and DV.

Comparison to other total effect measures. In the special case of a nominal outcome and a binary IV, our *total ME* approach matches the logic of a segregation or dissimilarity index for the unconditional relationship. For example, Duncan's index is a common measure to understand segregation (Duncan and Duncan 1955)

$$Duncan's\ index = \sum_{\lambda=1}^{\Lambda} \left| \frac{n_{IV=0|y=\lambda}}{n_{IV=0}} - \frac{n_{IV=1|y=\lambda}}{n_{IV=1}} \right| /2. \quad (30)$$

For example, we could use Duncan's index to understand how men and women (binary IV) are distributed across occupations (nominal outcome). In this example, a multinomial logit model regressing occupation on gender would produce a *total ME* equivalent to Duncan's index.

The *total ME* approach builds on this classic idea but allows for many more applications. Because the *total ME* is based on model predictions, it can be used to describe raw differences across groups or conditional differences. In addition, it can be used to understand effects not only for binary IVs but also for continuous and nominal/ordinal IVs. For example, Hällsten and Thaning (2018) use the logic of the Duncan index to calculate effects across the trimmed range of a continuous IV and then aggregate them as we suggest for the *total ME*. For most applications, we would use a +1 or +SD ME estimate for the *total ME* for a continuous IV except when comparing effects across IVs with different measurement levels (see the Comparing effect sizes across nominal/ordinal and continuous IVs section). We describe an approach for further expanding the *total ME* idea to nominal/ordinal IVs in the MEs for nominal or ordinal IVs in nominal and ordinal outcome models section.

Example of total ME summary measure to compare effects within a model. To illustrate our *total ME* summary measure for continuous and binary IVs, we examine predictors of self-rated health. We use GSS data from 2000 to 2021 and an ordinal outcome variable for *self-rated health* measured with four categories of poor, fair, good, or excellent. Although researchers often treat such outcomes as continuous in a linear regression, this can lead to biased results, as four categories are not a true continuum but are instead more accurately modeled as an ordinal or nominal outcome (Long and Freese 2014; Liddell and Kruschke 2018). We focus on the effects of *age* and *marital status* for this example; controls for *race-ethnicity*, *gender*, *parental status*, *family income*, and *education* are also included. We first tried an ordered logit but the model failed the Brant test, suggesting a violation of the ordinal model assumptions so we instead use a multinomial logit, which does not impose ordinality of the relationship between the outcome and the IVs (for details, see Long and Freese 2014). The *total ME* summary is calculated equivalently in ordinal and nominal outcome models so does not affect our calculation or interpretation.

To summarize the total effect of *age*, we first calculate MEs for a standard deviation increase in age for each outcome category. Figure 6 shows the results for age in the left panel. Aging a standard deviation (about 17 years) predicts a 2 percentage point increase in reporting poor health and a 3 percentage point increase in reporting fair health. In contrast, it predicts a 1 percentage point decrease in good health and a 4.5 percentage point decrease in excellent health.

The *total* ME_{age+SD} is simply the sum of the four individual MEs of age (taking absolute values) divided by two[18]

$$\begin{aligned}
Total\ ME = \Big[& \left|\eta_{poor}\left(age_i = age_i + SD_{age}, \mathbf{x}_{-k} = \mathbf{x}^*_{-k}\right) - \eta_{poor}\left(age_i = age_i, \mathbf{x}_{-k} = \mathbf{x}^*_{-k}\right)\right| \\
& + \left|\eta_{fair}\left(age_i = age_i + SD_{age}, \mathbf{x}_{-k} = \mathbf{x}^*_{-k}\right) - \eta_{fair}\left(age_i = age_i, \mathbf{x}_{-k} = \mathbf{x}^*_{-k}\right)\right| \\
& + \left|\eta_{good}\left(age_i = age_i + SD_{age}, \mathbf{x}_{-k} = \mathbf{x}^*_{-k}\right) - \eta_{good}\left(age_i = age_i, \mathbf{x}_{-k} = \mathbf{x}^*_{-k}\right)\right| \\
& + \left|\eta_{excellent}\left(age_i = age_i + SD_{age}, \mathbf{x}_{-k} = \mathbf{x}^*_{-k}\right) - \eta_{excellent}\left(age_i = age_i, \mathbf{x}_{-k} = \mathbf{x}^*_{-k}\right)\right| \Big] /2 \\
= & \frac{[0.021 + 0.033 + 0.009 + 0.045]}{2} = 0.054.
\end{aligned} \tag{31}$$

We find that the *total* ME_{age+SD} is 0.054 ($p < 0.01$), indicating that a SD increase in age affects self-rated health by about 5.4 percentage points in total. That is, the effect of a standard deviation increase in age is to shift about 5.4% of the sample to a different self-rated health category.

Marital status is a binary variable with married individuals coded 1. We calculate MEs for each outcome category and plot them as the right panel in Figure 6. We

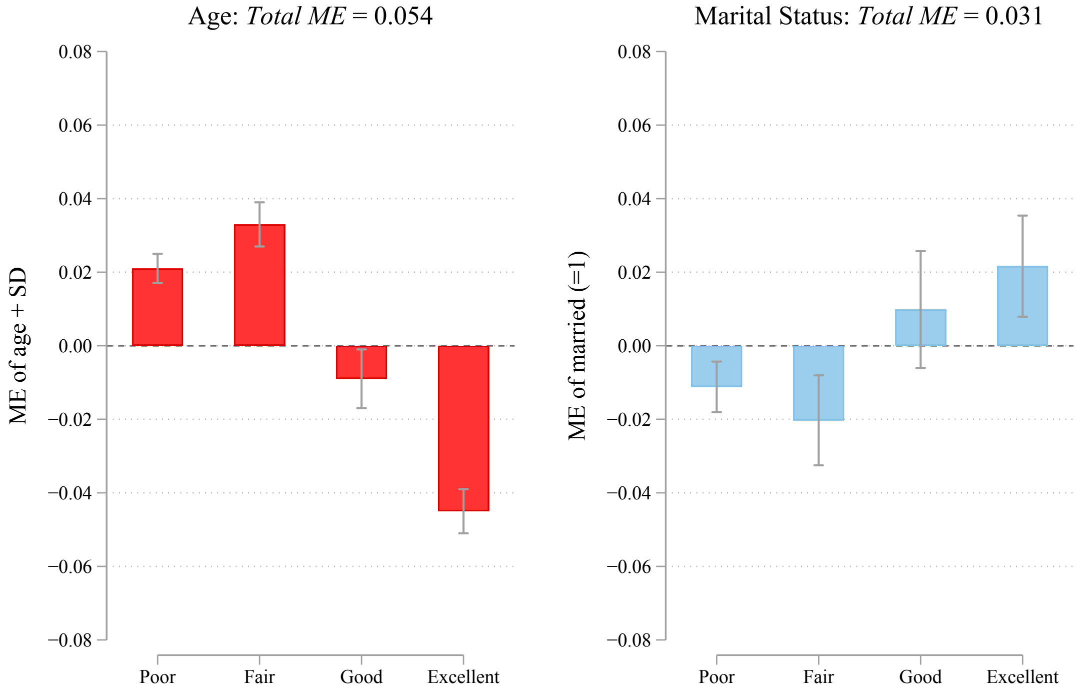

Figure 6: MEs of age (left panel) and marital status (right panel) on the probability of each self-rated health category.

find that married individuals report being in excellent health more often than nonmarried folks and are less likely to report poor or fair health. The *total ME* of marital status is 0.031 ($p < 0.01$), indicating that married and nonmarried individuals differ by about 3 percentage points total in terms of their self-rated health.

The *total ME* measure is useful for several purposes. First, it provides a single number and significance test of a variable's holistic effect on a nominal or ordinal outcome. Second, it is useful to understand the overall magnitude of an IVs effect. Third, it can be used to compare effect sizes across variables or across models/outcomes. For example, we recalculated the *total ME* of age using a +2 SD centered increase, as discussed in the Comparing effect sizes across nominal/ordinal and continuous IVs section for comparisons of continuous and binary IV effect sizes. We find that the $total\ ME_{age+2\ SD}$ is 0.103, which suggests that age's total effect on health is over three times larger than the total effect of marital status ($total\ ME_{age+2SD} - total\ ME_{married} = 0.103 - 0.031 \approx 0.071, p < 0.01$).

Simplifying tests of interaction with the total ME. Another application of the *total ME* is to simplify tests of interaction in nominal and ordinal outcome models. Current best practices suggest a test of interaction across each outcome category (Mize 2019), though this can be cumbersome and may answer a more nuanced question than necessary. To illustrate, we edit our self-rated health example from the last section by adding an interaction (product) term between *marital* and *parental status* (married = 1; parent = 1). Figure 7 plots the predicted probabilities of each self-rated health

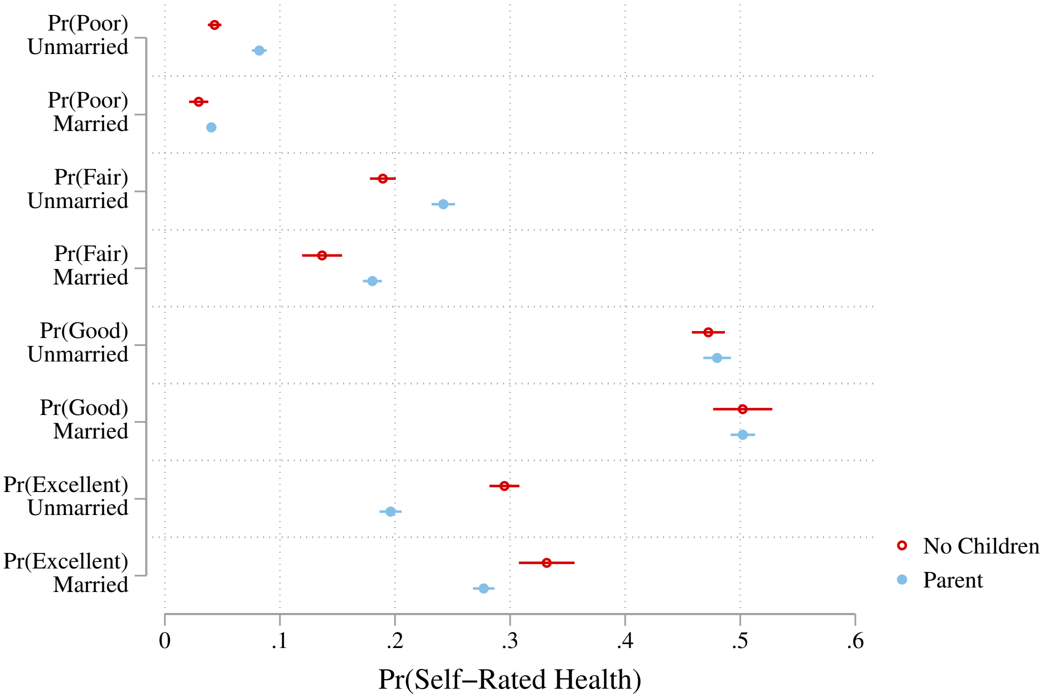

Figure 7: Predictions for self-rated health rating by marital and parental status.

outcome based on marital and parental status. The effect of parental status is represented by the gap between the two predictions in each row on the plot.

The left panel of Figure 8 visualizes the ME of parental status directly. The first row shows that for unmarried individuals, parents are more likely to report poor health than non-parents. The largest effect of parental status is for those who are unmarried on the probability of reporting excellent health, with these parents notably less likely to report excellent health.

The interaction effect can be tested as the second difference or equality of the two MEs of parental status—for the unmarried and the married—for each outcome (Mize 2019). The right panel of Figure 8 shows the second difference for each outcome category. Two of the four are significant, indicating that parental status has a larger effect for unmarried than for married individuals on the probability of reporting poor or excellent health (though in opposite directions). We find no evidence of interaction for the probability of reporting fair or good health.

Although the approach above provides a valid test of interaction, it provides an ambiguous answer given differences across outcome categories. In addition, it answers a more nuanced question than "*is the effect of parental status on health the same for married and unmarried individuals*?" If this broader question is of interest, *total MEs* can provide a holistic quantification of the effect of parental status to use in the test of interaction. In this case, we calculate two *total MEs* of parental status, one for the married and one for the unmarried, and then test their equality using a second difference test. The *total ME* of parental status is 0.017 for the married and 0.061 for the unmarried. The total effect of parental status is larger for the unmarried than

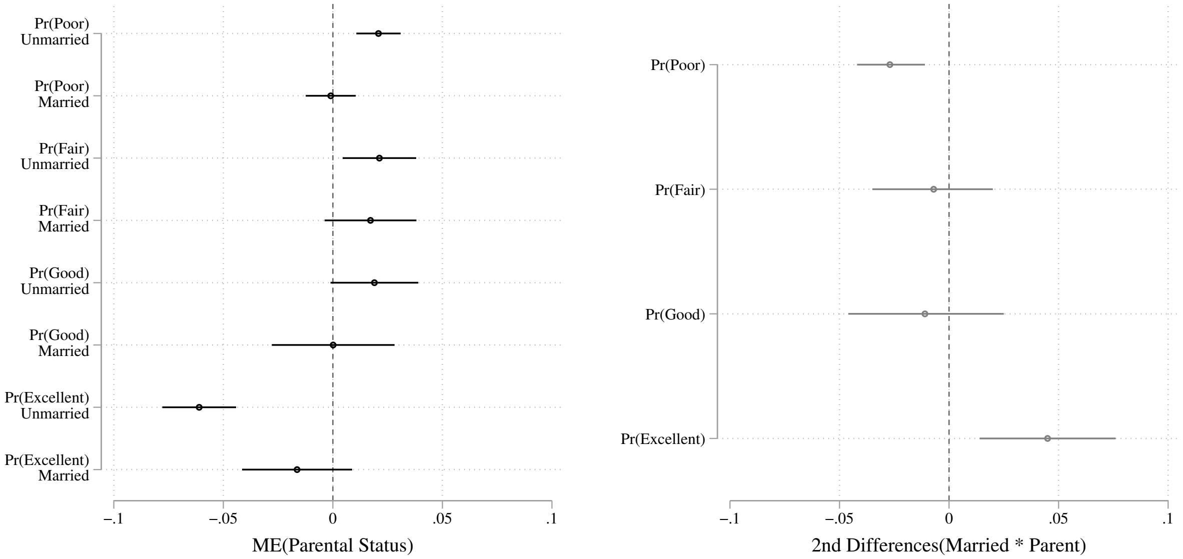

Figure 8: MEs of parental status on self-rated health rating, by marital status (left panel); and second differences for test of interaction between marital and parental status.

for the married (second difference of *total MEs* $= 0.061 - 0.017 = 0.044$, $p < 0.01$), indicating a significant interaction effect between parental and marital status.

Using total ME summary measures for tests of mediation/attenuation. Next, we use the *total ME* to compare the effect of a single focal IV across multiple models, such as in tests of mediation/attenuation. Often, researchers are interested in whether the overall effect of a variable changes or not across models, rather than whether the effect on each outcome category changes.

We tweak the example for self-rated health from the last section to test for mediation. In the first model, we include a *college degree* binary indicator as a focal IV along with demographic controls. In the second model, we add *family income* as a potential mediator of the effect of having a college degree on health.

Table 4 includes the results for each individual ME of college in the top panel. The third column of the table includes a test of the cross-model differences in the MEs of college. Here, we see that three of the four MEs are significantly different across the two models. However, none of these test the holistic question of whether the effect of college is attenuated or not after accounting for family income; the *total ME* answers this question.

As shown in the bottom panel of Table 4, the *total ME* of college in the first model is 0.159 and reduces to 0.118 in the second model with family income included. A test of the cross-model difference of *total MEs* is significant, suggesting that income partially explains the effect of college on health (cross-model difference $= 0.159 - 0.118 = 0.041$, $p < 0.01$); family income explains about 26 percent of the total effect of college on health.

Comparing effects across outcomes using the total ME. Another application of the total ME is to compare effects across different outcomes. For binary, nominal, and ordinal

Table 4: Marginal effects (MEs) and total MEs of college on self-rated health in base model and mediation model including family income.

(1)	(2) Base Model	(3) Mediation Model	(4)
	$ME_{college}$	$ME_{college}$	Difference
Pr(Poor)	-0.047^{*}	-0.033^{*}	0.015^{*}
Pr(Fair)	-0.112^{*}	-0.085^{*}	0.027^{*}
Pr(Good)	0.024^{*}	$0.021^{\dagger}$	-0.003
Pr(Excellent)	0.135^{*}	0.097^{*}	-0.038^{*}
	Total $ME_{college}$	*Total* $ME_{college}$	Difference
	0.159^{*}	0.118^{*}	0.041^{*}

Note: $^{\dagger}p < 0.05$, $^{*}p < 0.01$ in a two-tailed test.

outcome models, the prediction of interest is usually the predicted probability and thus provides a consistent metric facilitating comparison across these models. In each case, the predicted probabilities can always range from 0 to 1 and the *total MEs* will similarly have a potential range of 0 to 1.

To illustrate comparing across outcomes, we compare the total effect of gender ($woman = 1$) across three different opinion questions about religion included in the GSS. The first is a binary outcome assessing if the respondent thinks *atheists should be allowed free speech rights* (allow = 1). The second is an ordinal outcome assessing how much *confidence in organized religion* the respondent has (a great deal = 1; only some = 2; hardly any = 3). The third is a nominal outcome for respondent's *feelings about the bibl*e as the literal word of God (= 1), an inspired text (= 2), an ancient book of fables (= 3), or some other opinion (= 4). We fit a binary, ordinal, and multinomial logit for each outcome, respectively, including controls for *age*, *college*, and *race-ethnicity*.[19]

Table 5 presents the MEs of gender for the three models in column 2. Compared to men, women are less supportive of free speech rights for atheists, have more confidence in organized religion, and are more likely to see the bible as the literal word of God and not an ancient book of fables. Comparing individual effects across outcomes is problematic because there are different numbers of categories across each outcome and the categories are not comparable. However, the *total MEs* presented in the right column can be compared directly. Doing so suggests that gender has a similar sized effect on opinions about atheist free speech rights and confidence in organized religion (cross-model difference of *Total MEs* $= 0.056 - -0.054 = 0.002$, $p =$ n.s.). In contrast, gender has a larger effect on opinions about the bible than on the two other religion questions (both cross-model differences $p < 0.01$).

MEs for Nominal or Ordinal IVs in Nominal and Ordinal Outcome Models

Combining nominal or ordinal IVs with nominal or ordinal outcomes presents a challenge of an overwhelming number of predictions and MEs to examine. For a nominal IV with L categories and a nominal or ordinal outcome with Λ outcome

Table 5: Marginal effects (MEs) and total MEs of gender (*woman* = 1) on different opinions about religious issues.

(1)	(2) ME_{woman}	(3) *Total* ME_{woman}
Atheist free speech		
Pr(Allow)	-0.056^{*}	0.056^{*}
		(0.009)
Confidence in religion		
Pr(A Great Deal)	0.053^{*}	0.054^{*}
Pr(Only Some)	0.001	(0.009)
Pr(Hardly Any)	-0.054^{*}	
Feelings about Bible		
Pr(Word of God)	0.075^{*}	0.091^{*}
Pr(Inspired Word)	0.016	(0.010)
Pr(Ancient Book)	-0.082^{*}	
Pr(Other)	-0.009^{*}	

Note: Standard errors in parentheses. $^{\dagger}p < 0.05$, $^{*}p < 0.01$ in a two-tailed test.

categories, there are many MEs

$$\begin{aligned}&\textit{\#of MEs for nominal/ordinal IVs in nominal/ordinal DV models}\\&= \text{\# of } \textit{outcome categories } * \text{\# } \textit{of pairwise comparisons of the nominal/ordinal IV}\\&= \Lambda * \frac{L\left(L-1\right)}{2}.\end{aligned} \tag{32}$$

The MEs for a nominal or ordinal IV are still the pairwise comparisons of the predictions for each IV category but with the additional complication of calculating these separately for each nominal or ordinal outcome category λ. For example, for a three category IV

$$\begin{aligned}ME_{\lambda,\textit{nominal 1 vs. 2}} &= \eta_\lambda\left(x_k = 2, \mathbf{x}_{-k} = \mathbf{x}^{*}_{-k}\right) - \eta_\lambda\left(x_k = 1, \mathbf{x}_{-k} = \mathbf{x}^{*}_{-k}\right),\\ ME_{\lambda,\text{nominal 1 vs. 3}} &= \eta_\lambda\left(x_k = 3, \mathbf{x}_{-k} = \mathbf{x}^{*}_{-k}\right) - \eta_\lambda\left(x_k = 1, \mathbf{x}_{-k} = \mathbf{x}^{*}_{-k}\right),\\ ME_{\lambda,\textit{nominal 2 vs. 3}} &= \eta_\lambda\left(x_k = 3, \mathbf{x}_{-k} = \mathbf{x}^{*}_{-k}\right) - \eta_\lambda\left(x_k = 2, \mathbf{x}_{-k} = \mathbf{x}^{*}_{-k}\right),\end{aligned} \tag{33}$$

where we would need to repeat these calculations for each of the Λ outcome categories. This overwhelming number of MEs makes clear the need for summary measures.

ME inequality measures for nominal and ordinal outcomes. To start, we can apply the same methods derived in the Proposed ME inequality summary measures of nominal and ordinal IV effects section to calculate *ME inequality* measures separately for each outcome category. For example, the absolute average weighted ME inequality for the predicted probability of a specific λ outcome category is

$$\begin{aligned}&\textit{ME inequality}_\lambda \equiv \left|\textit{average weighted ME inequality}\right|_\lambda\\&= \sum_{a=1}^{L}\sum_{b=2,b>a}^{L}\frac{w_{a,b}}{L-1} * \left|\eta_\lambda\left(x_k = a, \mathbf{x}_{-k} = \mathbf{x}^{*}_{-k}\right) - \eta_\lambda\left(x_k = b, \mathbf{x}_{-k} = \mathbf{x}^{*}_{-k}\right)\right|.\end{aligned} \tag{34}$$

And we would repeat these calculations for all Λ outcome categories. This procedure will greatly simplify interpretation. For example, with a four-category outcome and a six-category IV, there are

$$\Lambda * \frac{L\,(L-1)}{2} = 4 * \frac{6\,(6-1)}{2} = 4 * 15 = 60, \tag{35}$$

60 MEs. In contrast, there are always only Λ *ME inequality*$_\lambda$ measures; in this case, four.

Total ME inequality summary measures for nominal/ordinal IVs. We can combine the *ME inequality* approach for summarizing nominal and ordinal IV effects with the *total ME* approach developed in the prior section to obtain a single summary measure. In this case, the *total ME inequality* for a nominal or ordinal IV is simply the sum of the *ME inequality*$_\lambda$ measures divided by two[20]

$$\textit{total ME inequality} = \sum_{\lambda=1}^{\Lambda} \textit{ME inequality}_\lambda / 2. \tag{36}$$

Where the measure is divided by two so the range of the *total ME inequality* statistic is 0 to →1. As with the *total ME* measures discussed in the MEs and total ME summary measures in nominal and ordinal outcome models section, the maximum value of *total ME inequality* is one, which represents the nominal IV groupings being completely distinct across the outcome categories and zero represents the nominal IV groupings all having the same outcomes.

Example of total ME inequality summary measure with nominal and ordinal IVs. We return to the model of self-rated health as an outcome with four categories for this example. The model includes a nominal IV *race-ethnicity* with four categories and an ordinal IV *educational attainment* with five categories.

Inequality statistics for race-ethnicity are shown in the top panel of Table 6. We find $\textit{ME inequality}_{\textit{race-ethnicity},\,\Pr(\textit{poor})} = 0.012$, which indicates that racial-ethnic groups differ by about one percentage point on average in terms of their ratings of their health as poor. In contrast, race-ethnicity explains more of the variation in ratings of health as excellent ($\textit{ME inequality}_{\textit{race-ethnicity},\,\Pr(\textit{excellent})} = 0.039$). To obtain a *total ME inequality* of race-ethnicity, we sum each of the individual *ME inequality* measures and divide by two

$$\begin{aligned} &\textit{Total ME inequality}_{\textit{race-ethnicity}} \\ &= \frac{\left[\textit{ME inequality}_{\textit{poor}} + \textit{ME inequality}_{\textit{fair}} + \textit{ME inequality}_{\textit{good}} + \textit{ME inequality}_{\textit{excellent}}\right]}{2} \\ &= \frac{.012 + .033 + .010 + .039}{2} \approx .047. \end{aligned} \tag{37}$$

The results suggest that racial-ethnic groups differ by a total of 4.7 percentage points in terms of their self-rated health ratings.

We next calculate inequality measures for educational attainment and then average them for a *total ME inequality* measure. Each statistic is shown in the bottom panel of Table 6. In general, we see larger effects of educational attainment

Table 6: Measures of *ME inequality* for each self-rated health outcome and *total ME inequality*, for effects of race-ethnicity and educational attainment.

	(1) *ME Inequality*	(2) *Total ME Inequality*
Race-ethnicity		
Poor	0.012*	0.047*
Fair	0.033*	(0.007)
Good	0.010	
Excellent	0.039*	
Educational attainment		
Poor	0.034*	0.116*
Fair	0.079*	(0.006)
Good	0.034*	
Excellent	0.085*	

Note: Standard error of *total ME inequality* measure in parentheses. $^{\dagger}p < 0.05$, $^{*}p < 0.01$ in a two-tailed test.

than for race-ethnicity. For example, the largest effect is on excellent health with $\mathit{ME\ inequality}_{educ,\,\mathrm{Pr}(excellent)} = 0.085$. In particular, the *total ME inequality* provides a single number to compare the holistic effects of the two IVs. The *total ME inequality* for education is 0.116 suggesting that educational attainment groups differ by almost 12 percentage points total in terms of their self-rated health ratings. This is larger than the effect for race-ethnicity, with racial-ethnic groups differing by a smaller 4.7 percentage points (difference in *total ME inequalities* $= 0.116 - 0.047 = 0.069$, $p < 0.01$).

Additional Considerations for Multi-category Outcomes

Although we have focused on the most common types of nominal and ordinal models here, our *total ME* approach is applicable to any multi-category outcome model. If there are multiple predictions of interest, for example, one for each outcome category, then our methods are applicable. For example, the generalized ordered logit model or continuation-ratio models for ordinal outcomes, or a discrete choice model for nominal outcomes, all produce predictions for each outcome category and our *total ME* summary would be straightforward to implement (Greene 2009; Williams 2016; Bauldry, Xu, and Fullerton 2018).

The ordered logit/probit model has an additional complication in that there are multiple potential predictions of interest. In this case, it may be of interest to make predictions in the latent variable or y^* metric with an ordinal model. If so, there would be a single prediction for y^*, regardless of how many outcome categories there are (Long 1997; Long and Freese 2014). In that case, the methods shown in the Proposed ME inequality summary measures of nominal and ordinal IV effects section for single outcome models can be used to quantify *ME inequality* in terms of y^* predictions. Although this method is straightforward to implement, we do caution that the logic of y^* predictions is based on untestable and sometimes questionable assumptions, so it is not usually our method of choice for understanding these models.

Software

Implementing the inequality and total effect summary measures shown here is possible but cumbersome with current software. The calculations are relatively straightforward—though sometimes tedious—but obtaining standard errors is more difficult. We provide example Stata and R code to reproduce all examples shown in this article at https://www.trentonmize.com/research.

In Stata, we use **margins** to calculate the predictions and MEs and **nlcom** to calculate the *ME inequality* and *total ME* statistics, both of which use the delta method to calculate standard errors (Dowd et al. 2014). For cross-model comparisons, we use the SUEST method, as detailed in Mize et al. (2019). In R, we use the **avg_comparisons** function of the **marginaleffects** package, which also uses the delta method.[21] We are not aware of a method in R to easily implement the cross-model comparisons of effects. However, as discussed in the Standard errors and significance tests section, bootstrapping is an alternative approach for standard error calculation that works well when the delta method is unfeasible or when there is not an existing package to automate calculations, and this has also been shown to be appropriate for cross-model comparisons of MEs (Mize et al. 2019). Our example R files use this approach.

We have written two companion Stata commands, **meinequality** and **totalme**, which implement all methods described in this article. These greatly simplify the implementation of the methods, usually requiring only a single line of code to calculate the statistics reported throughout the article. The commands can be downloaded at https://www.trentonmize.com/software. We also plan to add an *ME inequality* option to the Stata command **mecompare** to allow for non-standard comparisons of the inequality statistics within or across models (Mize et al. 2019).

Discussion and Conclusion

Much of social science can be described as the study of social groups. These groupings sometimes represent nominal groups, such as political parties or religions, and sometimes represent rank-ordered groupings, such as social classes or educational degrees. In either case, the methods for studying these nominal or ordinal variables are similar and result in challenges for interpretation. Given there are often many groupings, there are many ways to compare the groupings to understand an effect. In this article, we present new methods for summarizing effects for nominal and ordinal variables to better understand effects, simplify analyses, and enable new understandings.

For nominal or ordinal IVs, we propose measures of inequality as a summary measure of these variables' effects. Our basic idea is that a nominal or ordinal IV has an effect size in proportion to the amount of inequality it patterns. This simple idea matches the way many sociologists think about group differences and brings theory and methodology in close alignment (Allison 1978; Lundberg, Johnson, and Stewart 2021). And because we use MEs as the building blocks of our *ME inequality* measures, our method can be used in almost any models: linear, nonlinear, categorical, multilevel, longitudinal, and more.

For nominal or ordinal outcome variables, we propose a new total ME summary measure. The *total ME* quantifies a variable's holistic effect on an outcome, allowing for a single number statement of the impact of that variable. Our approach extends to IVs of any type and allows for easy comparisons of effect sizes across variables and across models.

Notes

[1] Another possibility is to calculate a joint test of significance for all the pairwise comparisons. In this case, the joint test would provide an answer as to whether there is any overall difference among the categories but not provide information about the size of the differences (for a discussion of a related approach, see the Joint tests and likelihood-ratio tests section).

[2] We use the language from the GSS question here. Specifically, the categories are "Black or African American" and "American Indian or Alaska Native."

[3] This is always true for linear models and true if predictions are averaged over observed values in non-linear/categorical models. Alternative calculations treat each group as equal in size in the calculation of the global mean. In most cases, this is not desirable, and the global mean should instead reflect the sizes of the groups in the data as the methods we present do. See a related discussion in the When to use weighted versus unweighted ME inequality section.

[4] The likelihood-ratio test statistic is approximately χ^2 distributed with degrees of freedom equal to the number of constraints imposed, allowing for the calculation of $p-$values.

[5] However, results tend to be all but identical when holding control variables at their means (for a marginal effect at the mean), so this is a perfectly acceptable alternative.

[6] Where the condition $b > a$ ensures that no redundant contrasts are included in the calculation (e.g., only categories 2 vs. 3 are compared and not also 3 vs. 2).

[7] With a binary outcome the *unweighted ME inequality* measure has a possible range of 0–0.667. The range cannot exceed 0.667 without weighting because all groups cannot be distinct. For example, with three groups and a binary outcome, one case with the largest inequality is $\Pr(y = 1|x_k = a) = 1$, $\Pr(y = 1|x_k = b) = 0$, and $\Pr(y = 1|x_k = c) = 0$. In this case, groups b and c are identical though as distinct from group a as possible and the *unweighted ME inequality* equals 0.667.

[8] This can be the proportion in the sample or a weighted estimate if survey weights are used in the corresponding regression model.

[9] This correction is made so that the sum of all the multipliers equals one.

[10] Unlike the *unweighted ME inequality*, the weighted *ME inequality* measure does not have a ceiling at 0.667 for binary outcomes and instead has a potential range of 0 to →1. This is another reason we preference this version of the statistic as the default.

[11] If bootstrapping is used we can either use the standard deviation of the estimates as the standard error estimate and use Equation 16 or we can alternatively use the distribution of the estimates themselves as percentile estimates of the confidence intervals. For example, the 2.5th and 97.5th percentiles of the estimates for 95 percent CIs (Efron and Tibshirani 1993).

[12] For binary outcome models, the effect size is often easier to intuitively understand given both the predictions and the *ME inequality* statistics are limited to a range of 0 to 1.

[13] In this example, each group is equally represented in the data (by experimental design) so the *ME inequality* and *unweighted ME inequality* are identical.

[14] Exploratory analyses show the effect of age diminishing at older ages but with the probability not approaching one suggesting a squared term and the age^2 coefficient is statistically significant at $p < 0.05$ supporting its inclusion in the model.

[15] There is evidence of overdispersion for the outcome ($\bar{y} = 2.645$, $var\,(y) = 8.392$). The overdispersion parameter ($p < 0.01$) in the negative binomial model indeed suggests the negative binomial over the Poisson model.

[16] To quantify the percent change in an effect size across models: $\left[1 - \frac{\textit{effect in model 2}}{\textit{effect in model 1}}\right] * 100$.

[17] For this example, collinearity is a potential concern as the three IVs likely explain overlapping variation.

[18] For continuous IVs, we recommend analysts specify the amount of change when reporting the *total ME* as the value will be dependent on the amount of change examined (e.g., $+1$ or $+$SD).

[19] The ordinal model passes the Brant test ($p = 0.312$), suggesting that it is appropriate for these data.

[20] An identical statistic can be derived from the *total MEs* for each pairwise comparison of nominal IV groupings. The *total ME* in this case is the sum of a given pairwise comparison (e.g., groups a vs. b) MEs across each outcome category, divided by two. These *total MEs* for each pairwise comparison can then be summed, weighted by their group sizes, which will result in the same statistic as the *total ME inequality*.

[21] The **avg_comparisons** function calculates the marginal effects (pairwise comparisons for a nominal/ordinal IV). The **hypothesis** option can then be used to calculate the *ME inequality* or the *total ME* statistic and its standard error. See the example R files for details.

References

Agresti, Alan. 2013. *Categorical Data Analysis.* 3rd ed. New York: Wiley.

Allison, Paul D. 1978. "Measures of Inequality." *American Sociological Review* 43(6):865–80. https://doi.org/10.2307/2094626.

Althouse, Andrew D. 2016. "Adjust for Multiple Comparisons? It's Not That Simple." *The Annals of Thoracic Surgery* 101(5):1644–45. https://doi.org/10.1016/j.athoracsur.2015.11.024.

Bauldry, Shawn, Jun Xu, and Andrew S. Fullerton. 2018. "Gencrm: A New Command for Generalized Continuation-Ratio Models." *The Stata Journal* 18(4):924–36, https://doi.org/10.1177/1536867X1801800410.

Blackwell, Matthew and Michael P. Olson. 2022. "Reducing Model Misspecification and Bias in the Estimation of Interactions." *Political Analysis* 30(4):495–514. https://doi.org/10.1017/pan.2021.19.

Bland, James R. and Amanda C. Cook. 2019. "Random Effects Probit and Logit: Understanding Predictions and Marginal Effects." *Applied Economics Letters* 26(2):116–23. https://doi.org/10.1080/13504851.2018.1441498.

Bouchet-Valat, Milan. 2022. "General Marginal-Free Association Indices for Contingency Tables: From the Altham Index to the Intrinsic Association Coefficient." *Sociological Methods & Research* 51(1):203–236., https://doi.org/10.1177/0049124119852389.

Breen, Richard, Kristian Bernt Karlson, and Anders Holm. 2018. "Interpreting and Understanding Logits, Probits, and Other Nonlinear Probability Models." *Annual Review of Sociology* 44:39–54. https://doi.org/10.1146/annurev-soc-073117-041429

Curran-Everett, Douglas. 2000. "Multiple Comparisons: Philosophies and Illustrations." *American Journal of Physiology-Regulatory, Integrative and Comparative Physiology* 279(1):R1–R8. https://doi.org/10.1152/ajpregu.2000.279.1.R1

Dagum, Camilo. 1998. "A New Approach to the Decomposition of the Gini Income Inequality Ratio." Pp. 47–63 in *Income Inequality, Poverty, and Economic Welfare*, edited by D. J. Slottje and B. Raj. Heidelberg: Physica-Verlag HD. https://doi.org/10.1007/978-3-642-51073-1_4

Dowd, Bryan E., William H. Greene, and Edward C. Norton. 2014. "Computation of Standard Errors." *Health Services Research* 49(2):731–50. https://doi.org/10.1111/1475-6773.12122

Duncan, Otis Dudley and Beverly Duncan. 1955. "A Methodological Analysis of Segregation Indexes." *American Sociological Review* 20(2):210. https://doi.org/10.2307/2088328

Efron, Bradley and R. J. Tibshirani. 1993. *An Introduction to the Bootstrap*. New York: Chapman and Hall. https://doi.org/10.1201/9780429246593

Fortin, Nicole, Thomas Lemieux, and Sergio Firpo. 2011. "Decomposition Methods in Economics." Pp. 1–102 in *Handbook of Labor Economics*, edited by O. Ashenfelter and D. Card. Elsevier. https://doi.org/10.1016/S0169-7218(11)00407-2

Freese, Jeremy and Sasha Johfre. 2022. "Binary Contrasts for Unordered Polytomous Regressors." *The Stata Journal* 22(1):125–33. https://doi.org/10.1177/1536867X221083900

García-Pérez, Miguel A. 2023. "Use and Misuse of Corrections for Multiple Testing." *Methods in Psychology* 8:100120. https://doi.org/10.1016/j.metip.2023.100120

Gelman, Andrew. 2008. "Scaling Regression Inputs by Dividing by Two Standard Deviations." *Statistics in Medicine* 27(15):2865–73. https://doi.org/10.1002/sim.3107

Gelman, Andrew, Jennifer Hill, and Masanao Yajima. 2012. "Why We (Usually) Don't Have to Worry About Multiple Comparisons." *Journal of Research on Educational Effectiveness* 5(2):189–211. https://doi.org/10.1080/19345747.2011.618213

Greene, William. 2009. "Discrete Choice Modeling." Pp. 473–556 in *Palgrave Handbook of Econometrics: Volume 2: Applied Econometrics*, edited by T. C. Mills and K. Patterson. London: Palgrave Macmillan UK. https://doi.org/10.1057/9780230244405_11

Greene, William. 2012. *Econometric Analysis*. 7th ed. Boston Munich: Prentice Hall.

Hällsten, Martin and Max Thaning. 2018. "Multiple Dimensions of Social Background and Horizontal Educational Attainment in Sweden." *Research in Social Stratification and Mobility* 56:40–52. https://doi.org/10.1016/j.rssm.2018.06.005

Heise, David R. 1972. "Employing Nominal Variables, Induced Variables, and Block Variables in Path Analyses." *Sociological Methods & Research* 1(2):147–73. https://doi.org/10.1177/004912417200100201

Johfre, Sasha Shen and Jeremy Freese. 2021. "Reconsidering the Reference Category." *Sociological Methodology* 51(2):253–69. https://doi.org/10.1177/0081175020982632

Karlson, Kristian B., Anders Holm, and Richard Breen. 2012. "Comparing Regression Coefficients Between Models Using Logit and Probit: A New Method." *Sociological Methodology* 42:286–313. https://doi.org/10.1177/0081175012444861

Lazic, Stanley E. 2024. "Why Multiple Hypothesis Test Corrections Provide Poor Control of False Positives in the Real World." *Psychological Methods*. https://doi.org/10.1037/met0000678

Liao, Tim Futing. 2022. “Individual Components of Three Inequality Measures for Analyzing Shapes of Inequality.” *Sociological Methods & Research* 51(3):1325–56. https://doi.org/10.1177/0049124119875961

Liddell, Torrin M. and John K. Kruschke. 2018. “Analyzing Ordinal Data with Metric Models: What Could Possibly Go Wrong?” *Journal of Experimental Social Psychology* 79(August):328–48. https://doi.org/10.1016/j.jesp.2018.08.009

Long, J. Scott. 1997. *Regression Models for Categorical and Limited Dependent Variables*. Thousand Oaks, CA: Sage.

Long, J. Scott and Jeremy Freese. 2014. *Regression Models for Categorical Dependent Variables Using Stata*. 2nd ed. College Station, Texas: Stata Press.

Long, J. Scott and Sarah A. Mustillo. 2021. “Using Predictions and Marginal Effects to Compare Groups in Regression Models for Binary Outcomes.” *Sociological Methods & Research* 50(3):1284–320. https://doi.org/10.1177/0049124118799374

Lundberg, Ian, Rebecca Johnson, and Brandon M. Stewart. 2021. “What Is Your Estimand? Defining the Target Quantity Connects Statistical Evidence to Theory.” *American Sociological Review* 86(3):532–65. https://doi.org/10.1177/00031224211004187

MacKinnon, David P., Chondra M. Lockwood, Jeanne M. Hoffman, Stephen G. West, and Virgil Sheets. 2002. “A Comparison of Methods to Test Mediation and Other Intervening Variable Effects.” *Psychological methods* 7(1):83–104. https://doi.org/10.1037/1082-989X.7.1.83

Melamed, David and Long Doan. 2023. *Applications of Regression for Categorical Outcomes Using R*. 1st ed. Boca Raton: Chapman and Hall/CRC. https://doi.org/10.1201/9781003029847-1

Mize, Trenton D. 2019. “Best Practices for Estimating, Interpreting, and Presenting Nonlinear Interaction Effects.” *Sociological Science* 6:81–117. https://doi.org/10.15195/v6.a4

Mize, Trenton D., Long Doan, and J. Scott Long. 2019. “A General Framework for Comparing Predictions and Marginal Effects across Models.” *Sociological Methodology* 49(1):152–89. https://doi.org/10.1177/0081175019852763

Mize, Trenton D. and Bing Han. Forthcoming. “Marginal Effects: Flexible Methods for Interpretation across Linear and Nonlinear Models.” in *Handbook on Data Modeling and Analysis*. Elgar.

Mize, Trenton D. and Bianca Manago. 2018a. “The Stereotype Content of Sexual Orientation.” *Social Currents* 5(5):458–78. https://doi.org/10.1177/2329496518761999

Mize, Trenton D. and Bianca Manago. 2018b. “Precarious Sexuality: How Men and Women Are Differentially Categorized for Similar Sexual Behavior.” *American Sociological Review* 83(2):305–30. https://doi.org/10.1177/0003122418759544

Mize, Trenton D. and Bianca Manago. 2022. “The Past, Present, and Future of Experimental Methods in the Social Sciences.” *Social Science Research* (108):102799, https://doi.org/10.1016/j.ssresearch.2022.102799

Mood, Carina. 2010. “Logistic Regression: Why We Cannot Do What We Think We Can Do, and What We Can Do about It.” *European Sociological Review* 26(1):67–82. https://doi.org/10.1093/esr/jcp006

Nguyen, Trang Quynh, Ian Schmid, and Elizabeth A. Stuart. 2021. “Clarifying Causal Mediation Analysis for the Applied Researcher: Defining Effects Based on What We Want to Learn.” *Psychological Methods* 26(2):255–71. https://doi.org/10.1037/met0000299

Perneger, Thomas V. 1998. “What’s Wrong with Bonferroni Adjustments.” *British Medical Journal* 316(7139):1236–38. https://doi.org/10.1136/bmj.316.7139.1236

Rothman, Kenneth J. 1990. “No Adjustments Are Needed for Multiple Comparisons.” *Epidemiology* 1(1):43. https://doi.org/10.1097/00001648-199001000-00010

Rubin, Mark. 2021. “When to Adjust Alpha during Multiple Testing: A Consideration of Disjunction, Conjunction, and Individual Testing.” *Synthese* 199(3–4):10969–11000. https://doi.org/10.1007/s11229-021-03276-4

Rubin, Mark. 2024. “Inconsistent Multiple Testing Corrections: The Fallacy of Using Family-Based Error Rates to Make Inferences about Individual Hypotheses.” *Methods in Psychology* 10:100140. https://doi.org/10.1016/j.metip.2024.100140

Schleuter, D., M. Daufresne, F. Massol, and C. Argillier. 2010. “A User’s Guide to Functional Diversity Indices.” *Ecological Monographs* 80(3):469–84. https://doi.org/10.1890/08-2225.1

Stevens, S. S. 1946. “On the Theory of Scales of Measurement.” *Science* 7:677–78.

Wasserstein, Ronald L. and Nicole A. Lazar. 2016. “The ASA Statement on p-Values: Context, Process, and Purpose.” *The American Statistician* 70(2):129–33. https://doi.org/10.1080/00031305.2016.1154108

Wessie, Jeroen. 1999. “Seemingly Unrelated Estimation and Cluster-Adjusted Sandwich Estimator.” *Stata Technical Bulletin* 9:231–48.

White, Michael J. 1986. “Segregation and Diversity Measures in Population Distribution.” *Population Index* 52(2):198–221. https://doi.org/10.2307/3644339

Whitt, Hugh P. 1986. “The Sheaf Coefficient: A Simplified and Expanded Approach.” *Social Science Research* 15(2):174–89. https://doi.org/10.1016/0049-089X(86)90014-1

Williams, R. 2012. “Using the Margins Command to Estimate and Interpret Adjusted Predictions and Marginal Effects.” *Stata Journal* 12(2):308–31. https://doi.org/10.1177/1536867X1201200209

Williams, Richard. 2016. “Understanding and Interpreting Generalized Ordered Logit Models.” *The Journal of Mathematical Sociology* 40(1):7–20. https://doi.org/10.1080/0022250X.2015.1112384

Williams, Richard A. and Abigail Jorgensen. 2022. “Comparing Logit & Probit Coefficients Between Nested Models.” *Social Science Research* 102802. https://doi.org/10.1016/j.ssresearch.2022.102802

Acknowledgments: We thank Shawn Bauldry and the audience at The Methodology Center at Purdue’s work-in-progress series for their helpful comments on this article. We also thank Jonathan Horowitz for a well-timed question that pushed us to further develop the methods for nominal and ordinal outcomes.

Trenton D. Mize: Departments of Sociology & Statistics (by courtesy) and The Methodology Center at Purdue University. E-mail: tmize@purdue.edu.
Bing Han: Department of Sociology, Purdue University. E-mail: han644@purdue.edu.