GOETHE UNIVERSITÄT FRANKFURT AM MAIN

VERSION: 11 MARCH 2022

A Generalized Ordered Logit Model to Accommodate Multiple Rating Scales

Markus Gangl
Goethe University Frankfurt

POLAR Working Paper #3

European Research Council
Established by the European Commission

POLAR
Polarization and Its Discontents: Does Rising Economic Inequality Undermine the Foundations of Liberal Societies?

www.polar-project.org

DISCLAIMER:

This work represents original research by the authors. The authors gratefully acknowledge funding from the European Research Council under the European Union's Horizon 2020 Programme (Grant agreement n° 833196-POLAR-ERC-2018-AdG). Neither the European Research Council nor the primary data collectors and the providers of the data used in this research bear any responsibility for the analysis and the conclusions of this paper.

We welcome comments and suggestions on this research, please contact the corresponding author for this working paper at:

mgangl@soz.uni-frankfurt.de

SUGGESTED CITATION:
Markus Gangl. 2022. A Generalized Ordered Logit Model to Accommodate Multiple Rating Scales. POLAR Working Paper #3. Frankfurt: Goethe University. Retrieved from www.polar-project.org, version dated 11 March 2022.

www.polar-project.org

A Generalized Ordered Logit Model to Accommodate Multiple Rating Scales

Markus Gangl
Goethe University Frankfurt am Main

POLAR Working Paper #3

Abstract
Rating scales are ubiquitous in the social sciences, yet may present practical difficulties when response formats change over time or vary across surveys. To allow researchers to pool rating data across alternative question formats, the paper provides a generalization of the ordered logit model that accommodates multiple scale formats in the measurement of a single latent rating construct. The resulting multi-scale ordered logit model shares the interpretation as well as the proportional odds (or parallel lines) assumption with the standard ordered logit model. A further extension to relax the proportional odds assumption in the multi-scale context is proposed, and substitution of the logit with other convenient link functions is equally straightforward. The utility of the model is illustrated from an empirical analysis of the determinants of confidence in democratic institutions that combines data from the European Social Survey, the General Social Survey, and the European and World Values Survey series.

Keywords
ordered logit model, rating scales, data harmonization, question format, survey data, data pooling, trend analyses, cross-nationally comparative analyses

Acknowledgements
I gratefully acknowledge the opportunity to draw on microdata from the European Social Survey, the European Values Study, the World Values Survey, and the General Social Sur-vey for the present research. Of course, none of the original data collectors nor the data archives providing the scientific use files are responsible for my use of the data, nor for any interpretation that I derive from the analyses. I furthermore gratefully acknowledge the generous funding provided by the European Research Council (Grant Agreement no. 833196 – POLAR – ERC-2018-ADG) for this research.

Introduction

Rating scales are one of the epitomes of survey research. It is the rare questionnaire indeed that would not incorporate some version of a Likert scale to tap into the intensity of respondents' agreement with some opinion or statement, ask respondents to rate their happiness or satisfaction with specific domains of their lives, or that supplies respondents with some rating scale to help them express degrees of emotional bonding with particular social groups or sentiments of trust and confidence in others and in societal institutions. The communal feature of all these various forms of rating scales in survey research is that researchers are interested in capturing respondents' location on some latent dimension. This dimension may often be conceptualized as a continuum of underlying attitudes or beliefs, yet the latent dimension is lacking any natural metric, and hence different locations on the continuum may only be approximated by providing verbal or numerical cues to respondents. These cues then imply an element of gradation – as when it may be presumed that a statement of "strongly agree" corresponds to a higher degree of affirmation than "agree", or the choice of a happiness score of 8 to convey a higher level of contentment than the choice of a score of 5 – but substantial ambiguity inevitably remains as to whether respondents are sharing a reasonably common understanding of the survey stimuli, or whether and when the number of response categories might be sufficiently large and the distance between them sufficiently evenly spaced in substantive terms to permit treating the empirical observations as satisfying a metric scaling level.

In empirical research, such fine-grained methodological discussions often also seem to stem from the fact that the statistical modeling of ordinal data is something like the poor relation of the standard linear or logit regression models that social scientists are extensively familiar with. Applied researchers may be aware of the ordered logit (or probit) model that is extending fundamental principles of categorical data analysis to the case of ordinally-scaled

dependent variables (see Long 1997:114-47, Cameron and Trivedi 2005:519-21, Agresti 2010, Wooldridge 2010:655-59, Greene 2012:824-32, Hosmer, Lemeshow and Sturdivant 2013:289-310) or of interval regression models that find application when metric data have been recorded in response categories (e.g., income brackets) rather than as point data in the original metric (e.g., Cameron and Trivedi 2005:532-35, Wooldridge 2010:783-85), but in practice still turn back to more basic models when being confronted with ordinal outcome data. It seems fair to say that most social scientists then routinely either seek to rationalize a metric interpretation, perhaps even explicitly acknowledging the approximation, in order to proceed with using standard linear regression on their data, or resort to identifying specific cutoff points on the ordinal outcome scale from either theoretical or empirical considerations, and then use a standard binary logit (or probit) model to analyze outcomes. And in many instances, these convenience techniques will in fact provide pragmatic statistical solutions that result in valid and empirically informative parameter estimates, certainly when judged against conventional inferential standards and targets in quantitative social science research, where researchers are typically focused on establishing the principal existence and direction of some hypothesized effect, rather than on evaluating any sharply quantified prediction on the magnitude or range of some particular effect on some well-specified metric that would be deemed observationally compatible with a researcher's theoretical model.

Yet even when often well-founded, the social scientist's statistical pragmatism may find its limits. With rating data, an important practical difficulty arises whenever response formats change over time or when they vary systematically across different surveys. In some such cases, it might be possible to devise data harmonization rules from, e.g., noting the equivalence of certain verbal cues ("agree", "fully agree") that are being provided to respondents to help anchor the response scale, and to then analyze the data specifically at those cutoff points or response thresholds that appear being consistently captured over time or across

surveys. In other cases, it is possible to achieve data pooling and joint estimation via suitable interval regression modeling, namely when the ordinal response scale may have been merely a data recording tool to either help respondents by providing them with informative categorizations of some underlying continuous metric, or to increase item response rates by permitting respondents to choose between outcome categories (e.g., income brackets) rather than having to disclose point information on the original metric scale; multiple response formats and variability in recording categories are in fact straightforward to handle in the interval regression routines of standard statistical packages. Yet in one important class of situations, the applied social scientist is lacking guidance and adequate statistical tools, and that is when response formats evidently differ, when verbal cues are incompatible or unavailable, and when the scaling level is genuinely ordinal at the point of data collection because the construct of interest is lacking any natural metric.

This of course is precisely the case of the typical Likert-type survey question that asks respondents to express their degree of agreement with some particular statement, to rate their happiness and satisfaction with different domains of life, or to state their degree of closeness and attachment to some community, political party or organization, where the same question has been asked repeatedly over time or in different countries and places, but where the precise response format of the question may have changed over time or may have varied across surveys and locations. In one survey, respondents may have been asked to rate their happiness on a scale between 1 and 7, the next survey provides a scale from 0 to 10, and yet another may use four response categories that have explicit verbal labels (“excellent”, “good”, “satisfactory” etc.) attached to them. Or respondents may have been asked to state their degree of confidence in some public institutions, yet the survey was initially using a 5-point Likert scale, then switched to an 11-point scale in some later wave, and yet another version of the questionnaire may have experimented with using a small number of verbal cues as

response categories. In these and other similar constellations that frequently arise in survey research, it would be attractive to be able to pool and analyze the rating data across question formats in order to either simply increase statistical power or, perhaps more importantly, to obtain the required leverage to address broader substantive questions on, for example, historical changes or cross-country differences in outcomes and processes that cannot be addressed by using the original data sources in isolation. Yet because the rating scales in question lack any natural metric, data pooling may seem impossible or may at least seem to require that researchers be prepared to accept an inevitable degree of arbitrariness in whatever data harmonization rules they may choose to adopt. As a more principled alternative, however, it is also possible to generalize the standard ordered logit model to accommodate the presence of multiple rating scales that capture a common latent construct or attitude, and to thereby resolve the apparent incommensurability of alternative response formats. In the remainder of this paper, I present and discuss the resulting multi-scale ordered logit model, and then illustrate its practical utility in an empirical analysis of the relationship between income inequality and citizens' trust in democratic institutions that draws on survey data from the European Social Survey (ESS), the General Social Survey (GSS), and the European and World Values Survey (EVS/WVS) series.

A Generalized Ordered Logit Model to Accommodate Multiple Rating Scales

Although various alternatives exist (see Fullerton 2009, Agresti 2010:44-117, Hosmer et al. 2013:289-310, Fullerton and Xu 2016), it is the so-called proportional odds model that is conventionally seen as the standard logit model for ordered outcome data (e.g., McCullagh 1980, Long 1997, Cameron and Trivedi 2005, Williams 2006, 2016). Besides retaining the straightforward interpretation and other features of the well-known logit model for binary outcome data, the proportional odds model rests on conceptual foundations that align with the typical

use of rating scales in social science surveys, and it will therefore also serve as the natural starting point for the proposed extension to a multi-scale version of the model that is capable of accommodating the presence of multiple rating scales in the data at hand. The proportional odds model itself may be conveniently written as the threshold model

$$Pr(Y_i > j) = \frac{\exp(\alpha_j + X_i\beta)}{1+\exp(\alpha_j + X_i\beta)} \quad \text{for } j = 1, 2, \dots, k-1 \tag{1}$$

to predict the probability that the observed response for respondent i is higher than the response category j on a rating (or otherwise ordered) scale consisting of k categories (Williams 2006, 2016).[1] This probability to cross threshold j is modeled as a standard logit function of a covariate vector X_i, a coefficient vector β, and a set of threshold parameters or cutpoints α_j. In this setup, the defining feature of the proportional odds model is that the structural component $X_i\beta$ to describe the association between respondent characteristics X_i and outcomes is assumed to be exactly the same at each of the cutpoints α_j defined by adjacent response categories. When writing out the model for the conditional odds of observing respondents with characteristics X_i in a response category higher than j as

$$\Omega(Y_i > j) = \frac{Pr(Y_i>j|X)}{Pr(Y_i\leq j|X)} = \frac{\frac{\exp(\alpha_j + X_i\beta)}{1+\exp(\alpha_j + X_i\beta)}}{\frac{1}{1+\exp(\alpha_j + X_i\beta)}} = \exp(\alpha_j + X_i\beta) \quad \text{for } j = 1, 2, \dots, k-1, \tag{2}$$

these turn out to be exactly proportional at each cutpoint location $\exp(\alpha_j)$. By implication, the effect of a change (or difference) in any particular covariate x may then be given by the simple expression for the odds ratio

$$OR_j = \frac{\Omega(Y_i>j|x, x_i+\Delta x)}{\Omega(Y_i>j|x, x_i)} = \frac{\exp(\alpha_j + (x_i+\Delta x)\beta)}{\exp(\alpha_j + x_i\beta)} = \exp(\Delta x \times \beta) \quad \text{for all } j, \tag{3}$$

[1] Various alternative, but mathematically equivalent parametrizations exist for the proportional odds ordered logit model (see Long 1997, Hosmer et al. 2013, Fullerton and Xu 2016). In line with the usage in Williams (2006, 2016), the threshold model formulation is adopted for ease of exposition here.

which is independent of the particular threshold j at which it is evaluated. In the proportional odds model, the same shift in a covariate x in other words implies the exact same proportionate shift in the odds of crossing a particular response threshold, irrespective of which specific response threshold j is being considered. The exact same features may also be described in terms of positing a linear model

$$y_i^* = \alpha + X_i\beta + \varepsilon \tag{4}$$

for a latent (continuous) variable y_i^* that is underlying the imperfect empirical observations available from respondents' choice of response category

$$y_i = j \text{ when } \alpha_{j-1} \leq y_i^* < \alpha_j \text{ for } j = 1, 2, \ldots, k,\ \alpha_0 = -\infty \text{ and } \alpha_k = +\infty. \tag{5}$$

The linear form of the structural model (4) implies that the same regression plane $X_i\beta$ gets shifted from cutpoint to cutpoint intercepts α_j, and hence the proportional odds model may also be characterized by pointing out the respective parallel regression (or parallel lines) assumption that it implies (see Long 1997:140-45, Wooldridge 2010:658-59, Williams 2006, 2016, Fullerton and Xu 2016:9-10, 21-24 for details).

As a modeling device, the parallel regression assumption has the powerful implication that the structural component $X_i\beta$ is independent of the precise response format of the rating scale employed to capture the underlying (continuous) dimension or attitude (see Long 1997:117-19 in particular). This feature is indeed the foundation for the proposed extension to a multi-scale ordered logit model, yet at the same time the parallel regression assumption also tends to be seen as overly restrictive (e.g., Williams 2006, 2016), and in practice amounts to the main reason why the ordered logit model tends to have a rather mixed reputation among social scientists. When confronted with textbook warnings along the lines of

> "My experience suggests that the parallel regression assumption is frequently violated … When the assumption of parallel regressions is rejected, alternative

> models should be considered that do not impose the constraint of parallel regressions." (Long 1997:145),
>
> "A key problem with the parallel-lines model is that its assumptions are often violated; it is common for one or more β's to differ across values of j; i.e., the parallel-lines model is overly restrictive." (Williams 2006:60) and
>
> "The use of an ordered logit model when its assumptions are violated creates a misleading impression of how the outcome and explanatory variables are related." (Williams 2016:11),

empirical researchers cannot be faulted for coming to the conclusion that, at least in its standard proportional odds form, the ordered logit model is hardly worth their attention.

Upon closer inspection, this widespread attitude as well as the implicit conflation of the two notions of "violated assumption" and "flawed model" it rests upon are quite misplaced, however. As can be seen from the latent variable formulation of the proportional odds model in equation (4), the effect of any covariate x implies nothing but a location shift along the underlying attitude or rating continuum. Equation (4), in other words, is thus nothing else than a categorical data analogue of the standard modeling assumption made by any researcher who decides to fit a linear OLS regression. Like in any OLS regression, the proportional odds model can thus be understood as a regression model for the central tendency of the latent outcome distribution, except that the outcome Y^* in this case is not directly observed, does not have any natural metric to guide the substantive interpretation, and that model identification requires the error term to be set to the logit distribution with a mean of zero and a variance of $\pi^2/3$ (or, alternatively, to the standard normal distribution if estimating an ordered probit model is being desired, see Long 1997, Cameron and Trivedi 2005, Wooldridge 2010). But when seen from this perspective, the actual meaning of the much-touted "violations" of the parallel regression assumptions becomes clearer, too: when the parallel regression

assumption is violated in an empirical analysis, this is a direct indication that the association between covariates x and the mean of the conditional outcome distribution does not exhaust the statistically detectable signals in the empirical data. The "violation" of the parallel regression assumption therefore merely implies that more can be learned from the data if the researcher decides to examine different parts of the outcome distribution instead of just focusing on its central tendency.

But this then is nothing like any inherent "failure" of the ordered logit model, and it clearly is something very different from the model being seen as "misrepresenting" how the outcome and explanatory variables are related. The proportional odds model (often) "misrepresents" the data in the exact same way that an OLS regression "misrepresents" it: both models focus on the central tendency of the outcome distribution and provide a linear regression model for it. Sometimes this is exactly what a researcher wishes for, because she may have a hypothesis to test on some average group difference in outcomes. In other cases, the researcher might have more encompassing descriptive interests or her hypothesis might be more complex because it relates (also) to a group difference in the variance of the outcome distribution or specifically to a group difference in one of the tails of the outcome distribution, and then standard OLS would be inadequate and the researcher would better turn to more appropriate (conditional) quantile regression models (e.g., Koenker 2005, Koenker, Chernozhukov, He et al. 2020), to (co)variance function regression techniques (e.g., Western and Bloome 2009, Bloome and Schrage 2021) or to other types of location-scale models (e.g., Hedeker and Nordgren 2013, Leckie, French, Charlton et al. 2014). But in neither case would anyone ever consider faulting the OLS regression model for principally "creat[ing] a misleading impression of how the outcome and explanatory variables are related" (Williams 2016:11). Instead, one would simply note that some inferential task is beyond the standard OLS regression, and then apply one of the readily available extensions of the basic model in

the empirical analysis. Tellingly, the surging interest in examining various types of heterogeneities in the relationships between purported causes and effects has been accompanied by a very visible increase in the use of quantile regression and related models to ascertain not just the association between a covariate and the mean outcome, but also group differences in the shape (or variance) of the entire outcome distribution (e.g., Cheng 2014, VanHeuvelen 2018a, b, Ebner, Kühhirt and Lersch 2020, Lersch, Schulz and Leckie 2020).

Seen in this light, what is usually perceived as a disadvantage and an "overly restrictive" nature of the ordered logit model (and its probit cousin), is actually a powerful feature of key interest to substantive research. For the specific case of ordinally recorded outcome data that can be understood as an imperfect measure of some underlying continuum, that is for the typical case of rating and other attitude data common in survey research, the proportional odds model is an elegant approach to model the central tendency of the outcome variable conditional on covariates and the assumption of a linear regression function. It is thus an analogue to the standard OLS regression, except that there also is a generalized ordered logit model to relax the parallel regression assumption when required (see Long 1997, Williams 2006, 2016, Fullerton and Xu 2016), and several formal statistical tests are indeed available to ascertain whether employing the generalized model may be indicated by some systematic signal in the empirical data (again, see Long 1997, Williams 2006, 2016, Fullerton and Xu 2016:109ff., and Brant 1990 for the well-known specification test). But the relation between the proportional odds model and the generalized ordered logit model certainly is not one between a "failure" and an "appropriate" model, but instead between a model that exclusively focuses on establishing group differences in the location of the conditional outcome distribution and an alternative model that addresses group differences in the mean and in the variance (i.e. in the shape as well as the location) of the conditional outcome distribution simultaneously. Clearly, the suitability of choosing one over the other is not a principal matter, but one

of research priorities, substantive questions and hypotheses – and unlike in the case of standard OLS regression, the ordered logit model even provides a unified framework for conducting either type of analysis.

Against this background, it may have become plausible why, despite much textbook criticism, the proportional odds model is nevertheless taken as the starting point for proposing a natural extension of the ordered logit model to a multi-scale setting. Indeed, it is precisely because of the assumption of parallel regressions and the associated equivalence of the model's structural component $X_i\beta$ across any and all response thresholds recorded in the actual survey instrument or data collection effort that such an extension is readily accomplished. Indeed, when the same structural component $X_i\beta$ can be assumed to govern reporting behavior (i.e. correctly describe the association between covariates X and observable outcomes) at each observable response category, then the exact question format of the rating scale is irrelevant and the empirical parameter estimates will not depend on the exact number or verbal cueing of response categories utilized in data collection (see Long 1997:117-19). But if that is the case, then the validity of the model will also not be affected if observations are being pooled across two or more surveys (or survey waves) employing somewhat different versions of a rating scale s to measure some latent outcome dimension Y^*. In other words, one may obtain a workhorse for situations where data pooling would be desirable for addressing substantive questions on, for example, over-time change or cross-country differences in attitude formation by specifying the multi-scale proportional odds model

$$Pr(Y_i > j_s | s_i = s) = \frac{\exp(\alpha_{j_s} + X_i\beta)}{1+\exp(\alpha_{j_s} + X_i\beta)} \quad (6)$$

$$\text{for } j_s = 1, 2, \dots, k_s - 1 \text{ and } s = 1,2, \dots, m.$$

As with other models from the family, this multi-scale model may be estimated by maximum likelihood and retains the straightforward interpretation as well as all other features of the standard ordered logit model, except for the fact that now several sets of cutpoints j_s, namely

one for each survey instrument $s = 1,2, \dots, m$, represent the observable response patterns Y, and that the parallel regression planes defined by $X_i\beta$ now shift outcomes along the cutpoints j_s of the particular scale type $s_i = s$ that respondent i happened to be confronted with. To fix ideas, I will now first demonstrate the principal utility of the multi-scale model in an illustrative analysis of citizens' trust in the national parliament that combines survey data from the European Social Survey, the General Social Survey, and the European and World Values Survey series, and then discuss a further generalization to relax the parallel regression assumption of the ordered logit model also in the multi-scale setting.

An Empirical Illustration: Explaining Confidence in Democratic Institutions

Citizens' trust in the institutions of government of course is a main pillar of any democratic order, and its study of consequently significant interest to the social sciences. Unsurprisingly, social science surveys regularly field questions that ask respondents to indicate how much confidence they have in specific institutions or branches of government, and provide them with different rating scale formats to express their degree of trust. For the purpose of providing an illustration of the multi-scale ordered logit model, one may note that high-quality and nationally representative surveys like the European Social Survey (ESS, European Social Survey 2018-2021), the General Social Survey (GSS, Smith, Davern, Freese et al. 2019), and the European and World Values Survey (EVS/WVS, Inglehart, Haerpfer, Moreno et al. 2014, European Values Study 1981-2017) series all contain respective items to solicit respondents' sense of trust in various democratic institutions. And especially when interested in exploring contextual determinants of citizens' trust in institutions, it would be attractive to be able to pool data across these various survey sources in order to increase empirical variation in institutional, societal or macroeconomic conditions, and to thereby enhance a study's analytical

leverage by fully exploiting the geographical or historical coverage of the available survey data.

Yet, unfortunately, as question formats differ significantly across surveys, it is far from self-evident how to best do that and how to achieve consistent data integration and valid data harmonization. In the EVS/WVS series, for example, respondents are asked to state whether they might have "a great deal" of trust, "quite a lot", "not very much" or "none at all", the ESS employs an 11-category rating scale, but only provides verbal anchors at either extreme of the scale (i.e. by linking the bottom category to the description of "no trust at all", and the top category to indicate "complete trust"), and the standard GSS question format allows respondents to distinguish whether they feel "a great deal", "only some", or "hardly any" confidence in public institutions (and still further variations exist, as the GSS has over the years occasionally tested alternative response formats). Confronted with this reality of technical variation in survey instruments, the ordinary response is that researchers either give up on the attempt of pooling the data entirely or that they are forced to bend over backwards in order to rationalize some particular data harmonization rule they may choose to adopt – for example, is the GSS category of having "only some" confidence referring to the same stimulus as the ESS/WVS category of "not very much" or might it not be closer to "quite a lot" of trust after all, and if so, which numbers from the 0-10 ESS rating scale might exactly reflect any of these verbal labels? – only to invariably find their statistical results to be questioned due to the inevitable arbitrariness involved in adopting any of several mildly plausible harmonization rules, and to perhaps eventually have their work accepted by peer reviewers and fellow researchers after demonstrating in extensive robustness checks that running the regression analyses under all the various harmonization rules that may claim some plausibility does not critically affect the main result a researcher wishes to report.

The purpose of the present paper is to introduce the multi-scale ordered logit model as a principled alternative to resolve the issue, and to permit standard regression modeling of the pooled data without requiring the analyst to resort to second-best data harmonization rules and all the smaller or larger degree of arbitrariness they may involve. To illustrate the working of the model more concretely, I now turn to an analysis of citizens' trust in the national parliament that seeks to characterize the empirical association between trust and respondent characteristics like gender, age and level of education on the one hand, as well as between trust and the macroeconomic environment described by a country's level of economic prosperity (i.e., GDP per capita) and its level of economic inequality (as indexed by the Gini coefficient of household equivalent disposable incomes) on the other. I conduct this analysis specifically for 2018, as biannual ESS data collection efforts are scheduled in even years and as the timing coincides with the 2017-2020 wave 7 of the WVS data collection and wave 5 of the EVS data collection that is coordinated with the larger WVS enterprise. Besides ensuring the principal availability of high-quality survey data for the period, the choice of focus on citizens' trust in 2018 is both slightly artificial and chosen for a strategic reason. Restricting the analysis to a single year of course is artificial insofar as more data is readily available in the original surveys, and as few analysts would therefore wish to limit their substantive research to some more restricted setting than the available survey data would easily afford. Yet over and above the practical goal of keeping the subsequent demonstration exercise reasonably parsimonious, the strategic element in this particular choice is that the data collection for the U.S. part of the WVS series had already happened in 2017. So, if it was the case that the (hypothetical) analyst truly wished for an international perspective on citizens' trust in the national parliaments in 2018 specifically, then she would need to bring in GSS data in order to keep the U.S. in the sample, as the GSS, like its ESS equivalent, is fielding its biannual data collection in even years. Yet doing so is inevitably upping the methodological stakes further,

as bringing in the GSS data implies that the researcher has to deal not just with two, but actually with three different response formats to measure the same dependent variable.

Different responses to tackling the challenge, and the differences in empirical conclusions they imply, may be inferred from Tables 1 and 2, which provide estimates from alternative regression models that may have been applied in practical research and that may be contrasted with the findings from a multi-scale ordered logit model that manages to unite all the available survey data under a single regression specification, despite the differences in response formats in the source data. Specifically, the parameter estimates for this multi-scale ordered logit model are the ones given as specification M1 in Table 1, and this model will be the focal point of comparison relative to other types of regression analyses that might plausibly have been conducted. True to purpose, model M1 is the regression specification that is utilizing the full sample of N=75,561 valid ESS, EVS, GSS, and WVS interviews that have been conducted during 2018, where respondents have provided a statement on their degree of trust in the national parliament as well as data on their gender, age and level of education, and where information on a country's level of economic prosperity and level of economic inequality could be obtained by merging data on previous-year (i.e., 2017) GDP per capita and on the previous-year Gini coefficient from the World Development Indicators (WDI, World Bank 2021) and the Standardized World Income Inequality Database (SWIID, Solt 2020), respectively. The resulting sample pools data from 44 countries and 52 national surveys, and the ESS, EVS and WVS source surveys each contribute roughly a third of the overall sample. More specifically, more than 26,000 cases from 21 European countries result from interviews that have been conducted in 2018 under the umbrella of the ESS round 9, another 23,000 interviews from 15 European countries stem from wave 5 of the EVS series, some 24,500 respondents have been surveyed in another 15 non-European countries in 2018 under the WVS wave 7, and a final sample of 1,500 U.S. respondents can be drawn from the 2018 GSS. Eight

European countries – namely, Austria, Germany, Estonia, France, Italy, Norway, Serbia, and the United Kingdom – contribute independent samples from the ESS as well as the EVS to the present analysis, which explains why the number of N=52 national surveys is exceeding the maximum number of N=44 countries.

TABLES 1+2 ABOUT HERE

To respect the hierarchical structure of the data, i.e. the fact that survey respondents are clustered within countries, I actually expand on the earlier presentation of the multi-scale ordered logit model by estimating its multilevel version

(7) $$Pr(Y_{ip} > j_s | s_{ic} = s) = \frac{\exp(\alpha_{j_s} + X_i\beta + u_p)}{1+\exp(\alpha_{j_s} + X_i\beta + u_p)}$$

$$\text{for } j_s = 1, 2, \dots, k_s - 1,\ s = 1,2, \dots, m, \text{ and } p = 1,2, \dots, q$$

for the present demonstration. Specifically, I am adding a standard $N(0, \sigma_u)$-distributed random effect u_p to the regression equation in order to account for contextual differences in levels of trust in the national parliament across *p*=52 country-survey waves and to correct the estimated standard errors for the clustering of respondents within these macro contexts, especially with a view towards ensuring valid statistical inference about any potential effects of macroeconomic context on citizens' trust in democratic institutions. In more substantively-minded applications, one might of course also be interested in expanding the model further by allowing for random slopes and by then examining sources of contextual heterogeneity in the coefficient vector β more systematically, but no such complications are being pursued here in order to keep the statistical model utterly parsimonious and focused on methodological essentials for the purpose of an illustrative demonstration. In much the same vein, none of the parameter estimates that are to be reported in the following should be seen as coming with any claims towards identifying any causal relationships in earnest. The remainder of this

paper will use the standard terminology of "effects" to describe parameter estimates for the statistical association between covariates X and the dependent variable Y, but it should be self-evident to the reader that the intended interpretation is purely associational and descriptive in the present context, and that further extensions as well as a more deliberate and theoretically grounded choice of controls would be required to seriously aim at anything more.

From these preludes, it is easy to summarize the substantive evidence from the multi-scale regression specification M1 as indicating that macroeconomic context as well as citizens' socio-demographics matter for trust in parliament. More specifically, the effect of GDP/capita on trust is positive but not statistically significant, yet high levels of economic inequality clearly depress citizens' trust in a core democratic institution like the national parliament. And on the microlevel, women on average are found to be slightly more skeptical of parliament than men, whereas higher levels of education lead to clearly higher levels of trust in the institutions of democratic governance. The age effects indicate a U-shaped pattern of association with democratic trust, with lowest levels of trust being found among citizens in their late forties, ceteris paribus.[2] And of course, as is true in any cross-sectional sample, what is reported as an age effect here is likely to reflect some mixture of true life-cycle and true cohort effects, but the fundamental identification problem at the heart of any age-period-cohort (APC) model of course prevents undertaking any empirically-grounded attempt to distinguish between and quantify the relative importance of either temporal source of political

[2] All quantitative covariates enter the model in grand mean-centered form. Respondents' mean age is close to 48 years in the analysis sample. Moreover, all substantive interpretation is kept deliberately colloquial and illustrative in the present context. The methodological literature on the proper interpretation of nonlinear probability models and on the closely related issues of comparing logit and probit coefficients across groups and model specifications and of interpreting interaction terms in nonlinear probability models is literally filling volumes, and is generally concluding that reporting average marginal effects on the probability scale is the preferred metric in all these cases and models. For the purposes of the present paper, however, it should be sufficient to note that the multi-scale ordered logit model of course also shares all respective features, possibilities and issues of interpretation with the whole logit family of regression models. For further background, I refer interested readers to Allison (1999), Williams (2009), Mood (2010), Karlson et al. (2012), Breen et al. (2018), Mize et al. (2019), and to the advanced textbook literature in the field.

trust – nor would any such attempt be required in the context of what is a purely descriptive and associational analysis done for demonstration purposes in the present paper.

Instead, the characteristic achievement of the multi-scale ordered logit model emerges when comparing model M1 against some more standard alternatives. Absent the multi-scale specification, it would of course have been possible to fit the standard, single-scale ordered logit model on the data, or at least on those parts of the sample that originate from the same source survey and therefore share the same question format in data collection. Estimates from respective specifications are provided as models M3-M7 in Table 1, each fitting a standard ordered logit model on data from one of the original survey sources or, in the last specification M7, on pooled EVS/WVS data that share the same response format for the political trust question. When eyeballing the parameter estimates across the different models, it is evident that some quite significant heterogeneity is apparent in the determinants of citizens' trust across surveys and localities – and that the estimates obtained in the multi-scale specification M1 provide something like the average over the different source surveys and over the whole sample of respondents. The effect of education, to take one example, is positive in the European data, but more so in the ESS sample than in the EVS one, but negative in the WVS sample, and quite negative in the GSS – and the multi-scale parameter estimate of $\beta = 0.076$ something like a reasonable estimate for the average effect of education on democratic trust for an overall sample that is dominated by European data. Similarly, at the macro level, there is evidence of a clear positive association between GDP per capita and democratic trust in the ESS data, a mildly positive, but non-significant association in the EVS sample, and a mildly negative, but non-significant association in the WVS sample – and then a mildly positive, but non-significant association of $\beta = 0.156$ as reported in the multi-scale estimate seems like a good estimate of the average relationship across all the countries in the sample. In addition, virtually all parameters are more precisely estimated, i.e. their standard errors are lower, in

the multi-scale specification M1 relative to its alternatives because the advantage of the larger (pooled) sample it is able to employ evidently outweighs any increase in variation that stems from pooling empirically heterogeneous data.

And this exact same pattern gets repeated if one was to compare the multi-scale estimates based on the two European sources (i.e., model M2) to those obtained from fitting standard ordered logit models on the ESS and EVS source data separately (i.e., to models M3 and M4) – and of course this is precisely the behavior that is to be expected from any regression model. It is very basic regression methodology to understand that any regression coefficient reflects the (weighted) average association between X and Y among the sample observations, and so fitting a regression model on some pooled data will inevitably result in parameter estimates that represent the weighted average of the corresponding coefficients from the series of identical models fitted on separate (and non-overlapping) partial datasets in isolation, and these parameter estimates will typically be more precisely estimated (i.e. exhibit lower standard errors) because of the larger sample brought to the task. And it is in this exact sense that the proposed multi-scale specification of the ordered logit model is shown to "work" as it should by the evidence of Table 1. It is a regression model that allows to pool rating data and that provides an estimate of the (weighted) average association between X and Y in the full sample, despite differences in response formats across source surveys.[3] The multi-scale model, in other words, substitutes a principled statistical model for any informal eyeballing that a researcher otherwise might wish to execute when trying to summarize rating scale evidence obtained from different samples and across different question formats.

[3] By the same token, the multi-scale specification of course allows to examine patterns of effect heterogeneity in greater detail. That some substantial degree of effect heterogeneity is present in the analysis sample is evident from Table 1 alone, and like in any standard (multilevel) regression model, one could expand on the simplistic specification adopted in this demonstration by, for example, introducing random coefficients and by then systematically considering cross-level interaction terms between the macroeconomic and the respondent-level covariates of the model. Even as this point is not specifically demonstrated in the empirical analysis here, it should be self-evident that increasing analysts' leverage to address and formally test for effect heterogeneity across contexts is another direct benefit of the proposed multi-scale specification relative to more traditional models.

The benefits of adopting this type of principled approach should also be self-evident when comparing the multi-scale ordered logit model to some more traditional convenience alternatives that are often being adopted to avoid the ordered logit model altogether. Table 2 provides some examples, and thereby helps further illustrate their downsides relative to the main multi-scale model (i.e., model M1 in Table 1). In applied research, social scientists often use standard OLS on rating data, and defend the practice by noting that substantive results more often than not tend to align with those of the ordered logit model. The same pattern is evident in the current analyses, as the linear regressions of M1 and M2 in Table 2 effectively mirror those of the corresponding ordered logit models M3 and M7 in Table 1 as far as the direction and statistical significance of the different effects are concerned.[4] But unlike with the ordered logit model, standard linear regression does not offer any constructive way forward when data pooling across different response formats may be desired. It is of course possible to adopt a rule-of-thumb harmonization protocol by, for example, distributing the four EVS/WVS response categories "evenly" across the 11-category ESS format in order to achieve data integration across these series, but whether that rule may have some empirical foundation or whether this amounts to a forced data pooling based on an entirely arbitrary methodological choice cannot adequately be decided.

Similarly, it would in principle be possible to define specific thresholds of the outcome variable that are of particular interest and then fit a standard binary logistic regression on the data, this approach would provide for a more intuitive interpretation of the resulting parameter estimates that is preferred by many social scientists over the linear model or the latent variable interpretation of the ordered logit model, but it would also not provide a constructive way forward to achieve valid data harmonization. Models M3-M6 are illustrations

[4] Given the GSS three-category response format, I do not present any separate linear regression modeling of the GSS data. Although estimation is certainly feasible, the required assumption of equidistance between categories seems to lack rather principal plausibility due to the small number of response categories in the GSS.

of the point, as these report the estimates from two logit model specifications that focus on the lower and the high end of the trust distribution, respectively, and that have each been fitted separately on the ESS and EVS/WVS data. These estimates on the one hand reflect similar substantive differences between the determinants of trust in the ESS and EVS/WVS samples and provide some empirical indications that associations between covariates and trust may indeed not be constant across the entire outcome distribution on the other – a topic to which I return in the next section –, but do not permit to answer the key question about the validity of pooling the data. Is, for the lower-tail models M3 and M4, a cutoff value of 4 on the ESS scale a good equivalent to respondents stating to have "not very much" trust on the EVS/WVS item, so that pooled analysis would be defensible? Is the ESS cutoff value of 8 about the same high level of political trust as expressed by EVS/WVS respondents who state having "a great deal of" trust in parliament?

Compared with these unanswerable questions, it may be instructive to consider how the multi-scale model addresses the comparability issue by effectively sidestepping it. Seen from the starting point of a latent continuous variable that is being imperfectly observed via the (ordered) categories of some particular rating scale employed in some specific survey, the multi-scale model is nothing but an extension of the standard ordered logit model that allows for the presence of multiple sets of scale location points in estimation, with one set of cutoff points corresponding to each type of question format. The empirical locations of the different cutoff points α_{j_s} are in fact being estimated as parameters of the model, and may therefore usefully be compared across response formats in order to assess what might be seen as the data harmonization rule that is implicit in the model and consistent with the empirical data. To continue the empirical example, Figure 1 provides the cutpoint locations α_{j_s} that have been estimated for the ESS, EVS/WVS and GSS scales in the main multi-scale model (M1 in Table 1), respectively; for easier reading, Figure 1 actually displays the inverted location

parameters $-\alpha_{j_s}$ as these correspond to a natural ordering of response categories in terms of increasing item "difficulty", i.e. to increasingly positive expressions of trust.[5]

FIGURE 1 ABOUT HERE

From these, it is readily apparent how the multi-scale specification is implicitly answering the earlier rhetorical questions: first, on the high end, the ESS cutoff value of 8 indeed seems to index pretty much the same intensity of trust as the verbal stimulus of "a great deal" of trust in the EVS/WVS surveys. But, second, on the low end, the ESS cutoff value of 4 does not seem to correspond to the EVS/WVS's verbal stimulus of having "not very much" trust. Instead, it rather is the ESS cutoff value of 3 that matches the EVS/WVS location of having "not very much" trust quite well, and Figure 1 then also suggests that the EVS/WVS category of expressing "quite a lot" of trust does not have its ready ESS equivalent, but is sitting somewhat uneasily between values 5 and 6 on the ESS scale.

But that said, it is also important not to mistake the evidence of Figure 1 for a suggestion of some substantive and empirically-grounded harmonization rule that might or that even should have been adopted by the researcher. Instead, the multi-scale model is better characterized as sidestepping the question of any substantive equivalence of response categories across different rating scales by combining a methodologically entirely relativist position on the "meaning" of any single response category with an additive model where the same structural component $X_i\beta$ to describe the associations between covariates X and outcomes Y gets shifted across successive cutpoint locations α_{j_s} along the distribution of the latent outcome.

[5] As is well known, two alternative formulations of the proportional odds model exist that are substantively entirely equivalent, but differ in the sign of the location parameters α (see Long 1997: 122-124). For didactical purposes, I prefer to build the ordered logit model from the positive probability of respondents crossing any particular response threshold, but which then suggests to invert location parameter estimates for easier interpretation. Alternatively, one could have built the model based on the probability of respondents staying below some response threshold with their recorded answer, and thereby arrive at the same set of parameters directly.

The best practical illustration for this point comes once again from considering the (strategically chosen) addition of the GSS question format as the third rating scale to be integrated into the full multi-scale model M1 in Table 1. Here, it does surprise the human researcher to see that the exact same verbal EVS/WVS and GSS stimuli of "a great deal" of trust are not being placed on par with each other in terms of their cutpoint locations α, but to see the EVS/WVS stimulus correspond to a cutoff value of 8 on the ESS scale, and the GSS category more to an ESS value of 9. Likewise, human researchers would probably not have equated the GSS stimulus of having "only some" trust with the ESS middle scale value of 5, and would have expected it to lie somewhere in between the EVS/WVS categories of having "not very much" and "quite a lot" of trust, but not quite to be almost the same as having "quite a lot" of trust as the empirical data seems to have it.

Yet of course, these estimated cutpoint locations α do not represent the outcome of any linguistic or substantive validation, but instead merely reflect the empirical reality of the conditional outcome distribution as observed via and anchored in some particular rating scale format. The fact that the locations of the GSS stimulus "a great deal" of trust and its EVS/WVS equivalent do not match, does not imply that the same words "mean" different things to respondents in different surveys and different countries in any substantive sense. It first and foremost means that the probability distributions differ in the sense that the share of GSS respondents who see themselves as having "a great deal" of trust is empirically smaller than the corresponding share of EVS/WVS respondents, of course averaging across all countries in the EVS/WVS sample and conditional on covariates in both cases. This relativist perspective sidesteps the question whether the difference is substantive or methodological, i.e. does not help decide whether GSS respondents differ from their EVS/WVS counterparts because Americans are truly showing less confidence in Congress than are citizens of other EVS/WVS countries in their national parliaments or because it truly is the case that the

stimulus of "a great deal" of trust may indeed convey different intensities of trust in the mind of U.S. respondents relative to respondents from other countries.

In the context of this current and slightly artificial analysis, it would thus not be possible to turn to the multi-scale ordered logit model to answer the substantive question whether U.S. citizens are less trustful of democratic institutions than the citizens of other countries, because the demonstration exercise has been strategically chosen to involve a perfect correlation between country and response format in the U.S. case. Hence, there is no extra degree of freedom available to estimate a country fixed effect and the cutoff locations α for the GSS response categories simultaneously, and all respective variation would in fact be attributed solely to the latter and hence become treated as a methodological nuisance parameter that is of little substantive interest to the analyst. But this, in turn, is not a bug, but indeed a feature and a decisive advantage of relying on the multi-scale model: a decision on the issue of what the various response categories may "mean" and how they may compare across countries and question formats is not required at all in order to move on and permit the social scientist to use the pooled sample and to evaluate the associations between covariates X and outcomes Y that are of genuine interest and a matter of theoretical reflection. The elegance of the model is that it sidesteps a problem that has been plaguing (comparative) survey researchers for decades and that may ultimately prove to be intractable in some respects, but that does not actually require a solution. And of course, it is a model that would allow to give an empirical answer to the question of whether U.S. citizens are less trustful of democratic institutions than the citizens of other countries or not. In any real-world analysis, one would of course not insist on using 2018 data exclusively, one would make sure to incorporate the 2017 U.S. sample from the WVS series, and one would thereby have broken what has been a perfect correlation between country and survey instrument in the artificial setup of the present exercise.

Relaxing the Parallel Regression Assumption in the Multi-Scale Ordered Logit Model

At this point, readers may agree with the perspective that it is possible to extend the ordered logit model to a multi-scale setting, while retaining a modeling framework that is well-known to social scientists and that affords flexible ways of interpreting the resulting parameter estimates either in terms of covariate effects on an underlying latent and continuous outcome or in terms of odds ratios or probability differentials of crossing specific response thresholds. At the same time, readers may likewise feel the multi-scale model to still be overly restrictive insofar as it of course also shares the critical parallel regression assumption with the standard ordered logit model. And even as some criticism of that assumption may itself be rather regarded as being based on a misapprehension, there is merit in the principal insistence on methodologies that permit researchers to adequately examine issues of dispersion and (treatment) effect heterogeneity over and above central tendencies of the outcome distribution and the association between covariates and conditional mean outcomes that are the mainstay of standard regression models including the proportional odds ordered logit model.

Respective interest in relaxing the parallel regression assumption may also be justified on purely empirical grounds, and in fact even in a somewhat artificial and restrictive setting like that of the present analysis. With dichotomous outcome measures to reflect, respectively, particularly high and low levels of political trust among citizens, a comparison of parameter estimates between ESS- and EVS/WVS-based logit models M3-M6 in Table 2 suggests that several relationships may in fact vary systematically across the outcome distribution, like the effect of gender and GDP per capita in the ESS data, or the effect of the Gini coefficient in the EVS/WVS sample. And in principle, it is actually straightforward to address such concerns and to relax the parallel regression assumption when required. As the multi-scale model is derivative of the standard ordered logit model, it also inherits the principal approach

toward its generalization. Specifically, and exactly as with the conventional model, the natural specification of a generalized multi-scale ordered logit model is

(8) $$Pr(Y_i > j_s | s_i = s) = \frac{\exp(\alpha_{j_s} + X_i \beta_{j_s})}{1 + \exp(\alpha_{j_s} + X_i \beta_{j_s})}$$

$$\text{for } j_s = 1, 2, \ldots, k_s - 1 \text{ and } s = 1,2, \ldots, m,$$

where the constant covariate vector β that generates the parallel regression planes in the standard model has been replaced by a cutpoint-specific covariate vector β_{j_s} that allows associations between covariates X and outcomes Y to freely vary at each observable response threshold (also see Fu 1998). The evident downside of this very general specification is that, as Williams (2006, 2016) correctly remarked, it may involve estimating many eventually superfluous parameters because some parameters may in fact be constant across the whole outcome distribution or at least show some more limited variation in certain parts of the distribution only, and then many parameters of the fully generalized model may in fact not be necessary as they are not (statistically significantly) different from each other. This consideration has brought Williams (2006, 2016) to proposing a partial proportional odds model that seeks to identify the exact minimal set of parameters that is required to describe the empirical structure of associations in a parsimonious and exhaustive way, while avoiding to estimate and report statistically superfluous coefficients.

While this procedure has evident statistical merit insofar as it seeks to exhaust the data signal by finding a maximally parsimonious model specification to fully capture and describe it, it is also possible to approach the issue of generalizing the ordered logit model less from a data-analytic and more from a subject-matter perspective. And without intending to deny the value of Williams' (2006, 2016) alternative, this will be the approach taken here. Then, from a subject-matter perspective, it often seems less relevant to be able to fully and efficiently characterize all systematic patterns that may be apparent in the empirical data, but generalizing from the standard proportional odds formulation seems warranted whenever researchers

may wish to evaluate hypotheses that extend beyond expectations about (conditional) group differences in average outcome levels. A typical case would seem to be that social scientists may harbor expectations about how some factor X would affect the shape of the outcome distribution over and above any upward or downward shift of the overall distribution that could be detected by examining conditional mean outcomes. Applied to the case at hand, one might reason that some covariates may be particularly relevant for protecting citizens from disenchantment with democratic institutions, and in such cases, one would expect to observe stronger associations between these particular covariates X and outcomes Y in the lower tail of the outcome distribution specifically, but weaker or perhaps even no statistical associations further up.

As one concrete example, adequate macroeconomic performance has often been considered a necessary condition of democratic legitimacy in classical works in political sociology or among students of the history of democratic societies (e.g., Lipset 1959, 1960, 2004). Translated into statistical terms, this could be read as indicating the expectation that macroeconomic conditions mostly affect the lower tail of the trust distribution, i.e. may be considered particularly decisive for determining whether someone accords at least some basic degree of confidence to the institutions of democratic governance, but may have fewer if any implications for whether someone may be expressing to trust some particular institution either "usually" or "almost all of the time." The substantive hypothesis in this case would imply a mean shift – trust in democratic institutions would be expected to be generally higher when macroeconomic environments are good than during a recession – but even more clearly it would involve an expectation about changes in the shape of the outcome distribution. Specifically, this consideration from classical political sociology would suggest that the variance of the outcome distribution increases during macroeconomic crises (or, as the flip side of the coin, that the trust distribution is relatively more compressed under normal times), and that

the increased variance comes about as the lower tail of the distribution fanning out because a certain share of the citizenry loses basic faith in the institutions of democratic governance under economic distress.

To test substantive hypotheses like these, neither the fully generalized ordered logit model nor Williams' (2006, 2016) partial proportional odds model would seem to fully meet the interests of substantively-minded social scientists. The fully generalized model clearly risks to provide excessive statistical detail, whereas the partial proportional odds model is setting statistical and data-analytic priorities rather than primarily substantive ones. Against that background, another type of generalization is to specify a generalized (and of course multiscale) ordered logit model in the form of

(9)
$$Pr(Y_i > j_s | s_i = s) = \frac{\exp(\alpha_{j_s} + X_i\beta_r)}{1+\exp(\alpha_{j_s} + X_i\beta_r)}$$

$$\text{for } j_s = 1, 2, \dots, k_s - 1,\ s = 1,2, \dots, m,$$

$$\text{and } r = \{j_s | \alpha_{j_s} \leq c_1\}, \{j_s | c_1 < \alpha_{j_s} \leq c_2\}, \dots, \{j_s | \alpha_{j_s} > c_t\}.$$

Here, a generalization from the parallel regression model has occurred insofar as β_r is no longer constant across the entire outcome distribution, but is allowed to vary systematically across tuples r of cutpoints j_s. When tuples are defined (across scale formats) by cutpoint locations α_{j_s} lying within some prespecified range $c_{r-1} < \alpha_{j_s} \leq c_r$ of the latent outcome distribution, it becomes possible to examine effect heterogeneity in β_r across different zones of the outcome distribution. One obvious example would be to define a single cutoff c towards the lower tail of the outcome distribution, and then evaluate whether covariate vectors $\beta_{\{j_s | \alpha_{j_s} \leq c\}}$ and $\beta_{\{j_s | \alpha_{j_s} > c\}}$ systematically differ in at least some of their elements in order to determine whether some covariates may indeed be particularly decisive at the lower end of the outcome distribution, i.e. to prevent citizens' disenchantment with democratic governance in the concrete example used in the present illustration.

The Empirical Example Continued:
Are there any Asymmetries in the Effects of Covariates on Citizens' Trust in Parliament?

To fix ideas, it seems straightforward to continue the earlier example, and to now examine whether some covariates may indeed show asymmetries in their association with citizens' trust in the national parliament, and if so, which covariates might be important to avoid citizens turning away from the institutions of democratic governance. As before, the actual model to be estimated will be the multilevel extension

(10)
$$Pr\left(Y_{ip} > j_s | s_i = s\right) = \frac{\exp(\alpha_{j_s} + X_i\beta_r + u_p)}{1+\exp(\alpha_{j_s} + X_i\beta_r + u_p)}$$

$$\text{for } j_s = 1, 2, \dots, k_s - 1,\ s = 1,2, \dots, m,\ p = 1,2, \dots, q,$$

$$\text{and } r = \{j_s | \alpha_{j_s} \leq c_1\}, \{j_s | c_1 < \alpha_{j_s} \leq c_2\}, \dots, \{j_s | \alpha_{j_s} > c_t\}$$

of the generalized multi-scale ordered logit model described in equation 9, and this model once again merely adds a context-level (country-survey wave) random effect u_p to reflect the hierarchical structure of the data. The resulting parameter estimates are provided in Table 3, which has the results from a simpler specification that contrasts lower-tail behavior to the associations observed at higher levels of the trust spectrum (model M1) and a second set of estimates from an expanded model (M2) that contrasts effects in the lower tail, the middle and the upper-tail of the distribution. In these generalized model specifications, I work with a cut-off of $c_{low} = logit(0.33)$ to define the lower tail of the distribution and $c_{high} = logit(0.75)$ to define the upper tail, i.e. I contrast associations in the lowest third of responses and, in model M2, in the highest quartile of responses to the remaining ones in the middle of the distribution. Given that the GSS response format comprises three response categories only, it is of course not possible to include the GSS data when estimating the more expansive model M2, because the desired contrast of effects across three areas of the outcome distribution of course requires that the outcome be observed on a rating scale that features at least four

response categories (i.e., at least three observable response thresholds between adjacent categories).

TABLE 3 ABOUT HERE

Empirically, it is self-evident from even the simplistic illustration of Table 3 why the standard proportional odds specification is often considered overly restrictive and substantively incomplete in the social sciences. Models 1 and 2 show clear evidence of effect heterogeneity in the associations between covariates and citizens' trust in the national parliament, in fact, it turns out that there is not one in these five rather basic covariates where the proportional odds (parallel regression) assumption may empirically be maintained. And as expected from the above stretch of the classical literature, macroeconomic context indeed matters mostly in the lower tail of the distribution, i.e. both prosperity and inequality are much more relevant in affecting citizens' basic level of trust in the institutions of democratic governance (relative to losing faith in them entirely) rather than for determining whether citizens' may be maintaining some dose of skepticism about institutions or may be seeing them as to deserve their full confidence. At the individual level, a similar observation apparently holds for citizens' level of education, although the effect does not become entirely non-significant, but retains a small positive association even at higher levels of trust. There also seems to be an interesting gender difference insofar as gender is affecting the shape of the outcome distribution much more strongly than its mean. Whereas gender differences (as in the evidence reported before) are relatively minor in the middle of the trust distribution, women are less likely to express extreme (i.e., either very low or very high) levels of political trust than men, and seem to particularly shy away from expressing to have full confidence in parliament. As this is evidence of a gender difference more in the shape than in the location of the outcome

distribution – with women's political evaluations being generally cast in more moderate terms than men's in this particular data, but not necessarily as more negative overall – it does indeed take a generalized ordered logit model specification to uncover the empirical regularity, and gender differences in democratic trust would hence appear as less consequential when looked at only through the lens of a standard proportional odds model.

Potentials and Pitfalls

Rating scales are ubiquitous in the social sciences, yet their widespread usage implies the equally ubiquitous problem of how to handle the situation that response formats might have changed over time or might systematically vary across surveys. To achieve data integration in such cases, social scientists conventionally seek to adopt some plausible data harmonization protocol that exhibits at least some fair degree of face validity. To help avoid the inevitable level of indeterminacy and arbitrariness involved in taking such methodological choices, the current paper has provided a straightforward generalization of the ordered logit model to a multi-scale setting as a principled alternative. As it accommodates multiple scale formats in the measurement of a single latent rating construct, the multi-scale ordered logit model permits the analyst to pool respective rating data even when it has been collected from different question formats, and it permits social scientists to work within the familiar and flexible statistical environment of logistic regression modeling as it inherits all standard features of the ordinary ordered logit model that it is descending from.

Even so, it is important to emphasize that the multi-scale model is not meant to suggest any canned solution to the intractable problem of determining the substantive equivalence of verbal or numerical stimuli across different surveys and survey instruments. Instead, the contribution of the multi-scale specification is to provide researchers with a statistical tool to effectively sidestep the comparability issue and to at the same time permit substantively

meaningful data analyses across alternative question formats that measure the same latent construct. The multi-scale model inevitably does rest on certain statistical assumptions as well, but these appear as rather benign and in fact are the exact same as implied in any conventional regression model for ordinal outcome data. The proposed model, like its standard ordered logit cousin, best fits the case of any rating scale that may be considered as a categorical recording device to approximate an underlying continuum. Conceptually, this seems an adequate way of thinking about many survey items to measure respondents' attitudes, beliefs or intensity of affiliation, but it is equally clear that other types of ordinal data exist that do not easily match this description. Likewise, even when conceptual foundations may appear adequate in principle, the proposed multi-scale model may still break down in practice and fail to produce meaningful results. When, for example, response categories happen to be very distinct and the overlap in response stimuli happens to be consequently slim across surveys, the model's implicitly relativist mode of data integration is likely to prove invalid at some point. But then, the purpose of the multi-scale specification is not and cannot be to transform imperfect data into ideal ones. It is to provide researchers with a way forward in an all-too common situation, and to exploit the imperfect data that we all rely on as best as may be under reasonably mild and practically defensible assumptions.

References

Agresti, Alan. 2010. *Analysis of Ordinal Categorical Data*. 2nd edition. Hoboken, NJ: Wiley.

Allison, Paul D. 1999. "Comparing Logit and Probit Coefficients across Groups." *Sociological Methods & Research* 28(2):186-208.

Bloome, Deirdre and Daniel Schrage. 2021. "Covariance Regression Models for Studying Treatment Effect Heterogeneity across One or More Outcomes: Understanding How Treatments Shape Inequality." *Sociological Methods & Research* 50(3):1034-72. doi: 10.1177/0049124119882449.

Brant, Rollin. 1990. "Assessing Proportionality in the Proportional Odds Model for Ordinal Logistic Regression." *Biometrics* 46(4):1171-78. doi: 10.2307/2532457.

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. doi: 10.1146/annurev-soc-073117-041429.

Cameron, A. Colin and Pravin K. Trivedi. 2005. *Microeconometrics: Methods and Applications*. Cambridge: Cambridge University Press.

Cheng, Siwei. 2014. "A Life Course Trajectory Framework for Understanding the Intracohort Pattern of Wage Inequality." *American Journal of Sociology* 120(3):633-700.

Ebner, Christian, Michael Kühhirt and Philipp Lersch. 2020. "Cohort Changes in the Level and Dispersion of Gender Ideology after German Reunification: Results from a Natural Experiment." *European Sociological Review* 36(5):814-28. doi: 10.1093/esr/jcaa015.

European Social Survey. 2018-2021. "European Social Survey Rounds 1-9 [Machine-Readable Data Files]." Bergen: NSD - Norwegian Centre for Research Data.

European Values Study. 1981-2017. "European Values Study 1981-2017 [Machine-Readable Data Files]." Cologne: GESIS Data Archive.

Fu, Vincent Kang. 1998. "Sg88: Estimating Generalized Ordered Logit Models." *Stata Technical Bulletin* 8:160-64.

Fullerton, Andrew S. 2009. "A Conceptual Framework for Ordered Logistic Regression Models." *Sociological Methods & Research* 38(2):306-47. doi: 10.1177/0049124109346162.

Fullerton, Andrew S. and Jun Xu. 2016. *Ordered Regression Models: Parallel, Partial, and Non-Parallel Alternatives*. Boca Raton, FL: Chapman and Hall/CRC.

Greene, William H. 2012. *Econometric Analysis*. 7th edition. Boston, MA: Pearson.

Hedeker, Donald and Rachel Nordgren. 2013. "Mixregls: A Program for Mixed-Effects Location Scale Analysis." *Journal of Statistical Software* 52(12):1-38. doi: 10.18637/jss.v052.i12.

Hosmer, David W., Stanley Lemeshow and Rodney X. Sturdivant. 2013. *Applied Logistic Regression*. 3rd edition. Hoboken, NJ: Wiley.

Inglehart, Ronald, Christian Haerpfer, A. Moreno, Christian Welzel, K. Kizilova, J. Diez-Medrano, M. Lagos, Pippa Norris, E. Ponarin and B. Puranen et al., eds. 2014. *World Values Survey: All Rounds - Country-Pooled Datafile Version*. Madrid: JD Systems Institute. (https://www.worldvaluessurvey.org/WVSDocumentationWVL.jsp).

Karlson, Kristian Bernt, Anders Holm and Richard Breen. 2012. "Comparing Regression Coefficients between Same-Sample Nested Models Using Logit and Probit: A New Method." *Sociological Methodology* 42:286-313. doi: 10.1177/0081175012444861.

Koenker, Roger. 2005. *Quantile Regression*. Cambridge: Cambridge University Press.
Koenker, Roger, Victor Chernozhukov, Xuming He and Limin Peng, eds. 2020. *Handbook of Quantile Regression*. Boca Raton, FL: CRC Press.
Leckie, George, Robert French, Chris Charlton and William Browne. 2014. "Modeling Heterogeneous Variance–Covariance Components in Two-Level Models." *Journal of Educational and Behavioral Statistics* 39(5):307-32.
Lersch, Philipp M., Wiebke Schulz and George Leckie. 2020. "The Variability of Occupational Attainment: How Prestige Trajectories Diversified within Birth Cohorts over the Twentieth Century." *American Sociological Review* 85(6):1084-116. doi: 10.1177/0003122420966324.
Lipset, Seymour M. 1959. "Some Social Requisites of Democracy: Economic Development and Political Legitimacy." *American Political Science Review* 53(1):69-105.
Lipset, Seymour M. 1960. *Political Man: The Social Bases of Politics*. Garden City, NY: Doubleday.
Lipset, Seymour M. 2004. *The Democratic Century*. Norman, OK: University of Oklahoma Press.
Long, J. Scott. 1997. *Regression Models for Categorical and Limited Dependent Variables*. Thousand Oaks, CA: Sage.
McCullagh, Peter. 1980. "Regression Models for Ordinal Data." *Journal of the Royal Statistical Society, Series B* 42(2):109-42. doi: 10.1111/j.2517-6161.1980.tb01109.x.
Mize, Trenton D., Long Doan and J. Scott Long. 2019. "A General Framework for Comparing Predictions and Marginal Effects across Models." *Sociological Methodology* 49:152-89. doi: 10.1177/0081175019852763.
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. doi: 10.1093/esr/jcp006.
Smith, Tom W., Michael Davern, Jeremy Freese and Stephen L. Morgan. 2019. "General Social Surveys, 1972-2018 [Machine-Readable Data File]." Chicago, IL: NORC.
Solt, Frederick. 2020. "Measuring Income Inequality across Countries and over Time: The Standardized World Income Inequality Database." *Social Science Quarterly* 101(3):1183-99.
VanHeuvelen, Tom. 2018a. "Within-Group Earnings Inequality in Cross-National Perspective." *European Sociological Review* 34(3):286-303. doi: 10.1093/esr/jcy011.
VanHeuvelen, Tom. 2018b. "Recovering the Missing Middle: A Mesocomparative Analysis of within-Group Inequality, 1970–2011." *American Journal of Sociology* 123(4):1064-116. doi: 10.1086/695640.
Western, Bruce and Deirdre Bloome. 2009. "Variance Function Regressions for Studying Inequality." *Sociological Methodology* 39:293-326. doi: 10.1111/j.1467-9531.2009.01222.x.
Williams, Richard. 2006. "Generalized Ordered Logit/Partial Proportional Odds Models for Ordinal Dependent Variables." *Stata Journal* 6(1):58-82.
Williams, Richard. 2009. "Using Heterogeneous Choice Models to Compare Logit and Probit Coefficients across Groups." *Sociological Methods & Research* 37(4):531-59. doi: 10.1177/0049124109335735.
Williams, Richard. 2016. "Understanding and Interpreting Generalized Ordered Logit Models." *Journal of Mathematical Sociology* 40(1):7-20. doi: 10.1080/0022250X.2015.1112384.
Wooldridge, Jeffrey M. 2010. *Econometric Analysis of Cross Section and Panel Data*. 2nd edition. Cambridge, MA: MIT Press.

World Bank. 2021, "World Development Indicators", Washington, D.C.: World Bank. (https://data.worldbank.org/).

Tables and Figures

TABLE 1
Macroeconomic and demographic determinants of trust in the national parliament under alternative ordered logit model specifications

	M1 Multi-scale Full sample	M2 Multi-scale ESS/EVS sample	M3 Standard ESS sample	M4 Standard EVS sample	M5 Standard GSS sample	M6 Standard WVS sample	M7 Standard EVS/WVS sample
Log_2 GDP/capita	0.156	0.302***	0.472***	0.177†	-	-0.175	0.018
	(0.107)	(0.081)	(0.135)	(0.085)		(0.239)	(0.137)
Gini	-0.046*	-0.079***	-0.064	-0.091***	-	-0.099*	-0.053†
	(0.023)	(0.024)	(0.039)	(0.026)		(0.046)	(0.027)
Female	-0.025*	-0.044**	-0.074***	0.016	0.062	-0.009	0.011
	(0.012)	(0.016)	(0.021)	(0.024)	(0.102)	(0.020)	(0.015)
Level of education	0.076***	0.157***	0.203***	0.091***	-0.101*	-0.046***	0.010†
	(0.005)	(0.007)	(0.009)	(0.010)	(0.048)	(0.008)	(0.006)
Age/10	0.009*	0.004	-0.034***	0.078***	-0.163***	0.005	0.047***
	(0.004)	(0.005)	(0.006)	(0.007)	(0.029)	(0.007)	(0.005)
Age/10 squared	0.030***	0.044***	0.054***	0.029***	0.043**	0.014***	0.022***
	(0.002)	(0.002)	(0.003)	(0.004)	(0.016)	(0.004)	(0.002)
N respondents	75,561	49,454	26,201	23,253	1,518	24,589	47,842
N survey waves	52	36	21	15	1	15	30
N countries	44	28	21	15	1	15	30

Notes: Selected parameter estimates from multilevel model specifications. All specifications nest respondents within country-survey waves. Standard errors in parentheses, statistical significance levels indicated at † p<.10, * p<.05, ** p<.01, and *** p<.001.

Sources: European Social Survey, European Values Study, General Social Survey, World Values Study, interviews conducted during 2018; macroeconomic indicators from the Standardized World Income Inequality Database and World Development Indicators

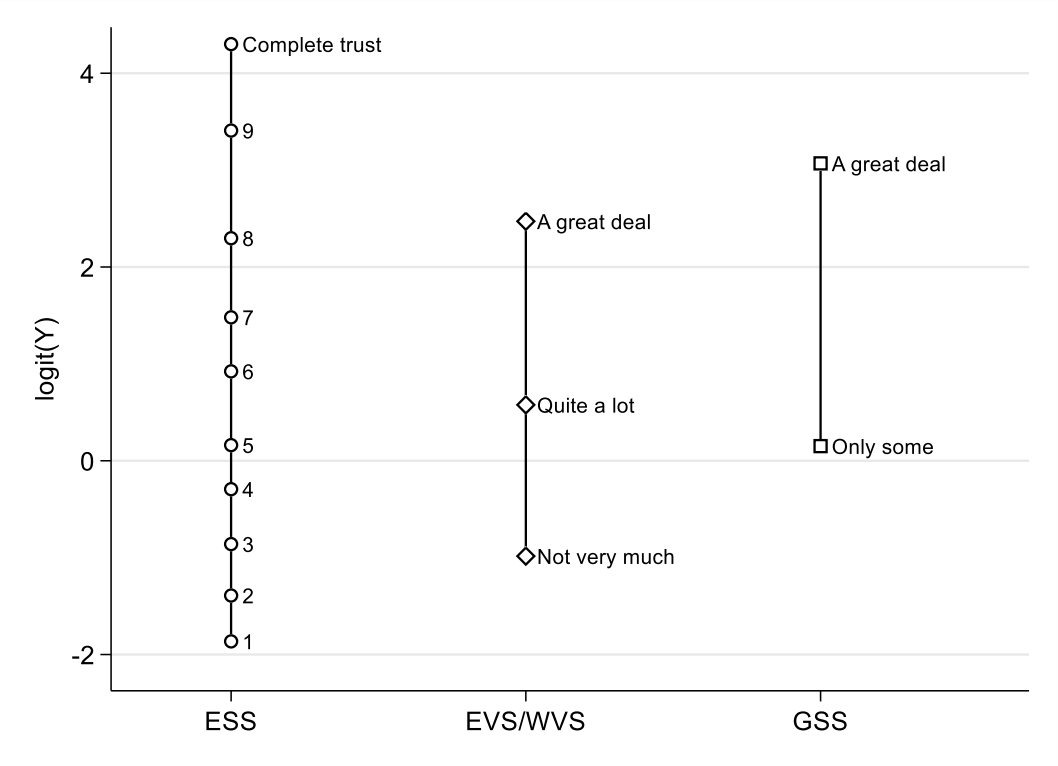
TABLE 2
Macroeconomic and demographic determinants of trust in the national parliament under alternative data pooling and standard regression specifications

	M1 OLS ESS, 11-point scale	M2 OLS EVS/WVS 4-point scale	M3 Logit ESS, low cutpoint	M4 Logit EVS/WVS, low cutpoint	M5 Logit ESS, high cutpoint	M6 Logit EVS/WVS, high cutpoint
Log_2 GDP/capita	0.657***	0.011	0.501***	-0.174	0.281	-0.212
	(0.183)	(0.065)	(0.137)	(0.124)	(0.175)	(0.206)
Gini	-0.088†	-0.026*	-0.056	0.061*	-0.056	-0.042
	(0.053)	(0.013)	(0.039)	(0.024)	(0.051)	(0.040)
Female	-0.104***	0.006	0.011	-0.059**	-0.199***	-0.101**
	(0.029)	(0.007)	(0.028)	(0.021)	(0.035)	(0.037)
Level of education	0.281***	0.005†	0.202***	-0.034***	0.174***	-0.063***
	(0.013)	(0.003)	(0.013)	(0.009)	(0.015)	(0.015)
Age/10	-0.048***	0.023***	-0.044***	-0.037***	-0.005	0.068***
	(0.008)	(0.002)	(0.008)	(0.007)	(0.009)	(0.011)
Age/10 squared	0.075***	0.011***	0.063***	-0.023***	0.048***	0.031***
	(0.004)	(0.001)	(0.004)	(0.004)	(0.005)	(0.006)
N respondents	26,201	47,842	26,201	47,842	26,201	47,842
N survey waves	21	30	21	30	21	30
N countries	21	30	21	30	21	30

Notes: Selected parameter estimates from multilevel model specifications. All specifications nest respondents within country-survey waves. Standard errors in parentheses, statistical significance levels indicated at † p<.10, * p<.05, ** p<.01, and *** p<.001.

Sources: European Social Survey, European Values Study, World Values Study, interviews conducted during 2018; macroeconomic indicators from the Standardized World Income Inequality Database and World Development Indicators

FIGURE 1
Estimated cutpoint locations for the ESS, EVS/WVS and GSS response formats to express trust in the national parliament



Notes: Inverted cutpoint estimates $-\alpha_{j_s}$ from model specification M1 in Table 1 (multi-scale ordered logit model, full ESS-EVS-GSS-WVS sample).

TABLE 3
Macroeconomic and demographic determinants of trust in the national parliament
under generalized multi-scale ordered logit model specifications

	M1 Full ESS-EVS-GSS-WVS sample			M2 ESS-EVS-WVS sample				
	Lower tail effects	Diff.	Main effects	Lower-tail effects	Diff.	Main effects	Diff.	Upper-tail effects
Log_2 GDP/capita	0.282**	***	0.095	0.282*	***	0.138	***	-0.009
	(0.107)		(0.107)	(0.109)		(0.109)		(0.108)
Gini	-0.057*	***	-0.039†	-0.057*	***	-0.044†	***	-0.024
	(0.023)		(0.023)	(0.023)		(0.023)		(0.023)
Female	0.039*	***	-0.057***	0.039†	**	-0.021	***	-0.138***
	(0.017)		(0.014)	(0.018)		(0.015)		(0.020)
Level of education	0.132***	***	0.051***	0.136***	***	0.066***	***	0.030***
	(0.007)		(0.006)	(0.007)		(0.006)		(0.008)
Age/10	0.003	*	0.014**	0.008		0.009†	***	0.030***
	(0.005)		(0.001)	(0.005)		(0.004)		(0.006)
Age/10 squared	0.034***	*	0.027***	0.034***	†	0.028***		0.026***
	(0.003)		(0.002)	(0.003)		(0.002)		(0.003)
N respondents		75,561				74,043		
N survey waves		52				51		
N countries		44				43		

Notes: Selected parameter estimates from multilevel model specifications; lower-tail cutpoints defined by $c_{low} \leq \text{logit}(0.33)$, upper-tail cutpoints by $c_{high} > \text{logit}(0.75)$. All specifications nest respondents within country-survey waves. Standard errors in parentheses, Wald tests for equivalence of parameter estimates across segments of the outcome distribution, statistical significance levels indicated at † p<.10, * p<.05, ** p<.01, and *** p<.001.

Sources: European Social Survey, European Values Study, General Social Survey, World Values Study, interviews conducted during 2018; macroeconomic indicators from the Standardized World Income Inequality Database and World Development Indicators