Latent-variable modeling of ordinal outcomes in language data analysis

Lukas Sönning, Manfred Krug, Fabian Vetter, Timo Schmid, Anne Leucht & Paul Messer

University of Bamberg

Abstract. In empirical work, ordinal variables are typically analyzed using means based on numeric scores assigned to categories. While this strategy has met with justified criticism in the methodological literature, it also generates simple and informative data summaries, a standard often not met by statistically more adequate procedures. Motivated by a survey of how ordered variables are dealt with in language research, we draw attention to an un(der)used latent-variable approach to ordinal data modeling, an alternative perspective on the most widely used form of ordered regression, the cumulative model. Since the latent-variable approach is not mentioned in statistical textbooks for linguists and does not feature in any of the studies in our survey, we believe it is worthwhile to promote its benefits. To this end, we draw on questionnaire-based preference ratings by speakers of Maltese English, who indicated on a 5-point scale which of two synonymous expressions (e.g. *package-parcel*) they (tend to) use. We demonstrate that a latent-variable analysis affords nuanced and interpretable data summaries that can be visualized effectively, while at the same time avoiding limitations inherent in mean response models (e.g. distortions induced by floor and ceiling effects). The online supplementary materials include a tutorial for its implementation in R.

1. Introduction

Ordinal variables consist of a set of ordered categories, where the ordering reflects a larger or smaller amount of what is being measured. Their midway position between continuous and nominal variables has given rise to a variety of analysis strategies in empirical work. Most commonly, ordered outcomes are translated into numeric scores and then treated in the same way as continuous variables. We will follow Agresti (2010: 137) and use the term *mean response model* (MRM) to refer to this form of analysis. It involves the calculation of means (and standard deviations), or the use of ordinary least-squares regression (or ANOVA). While MRMs have drawn justified criticism in the statistical literature (e.g. Long 1997: 116–119; Agresti 2010: 5–8), one of their key advantages is the informative and interpretable output they produce – they allow us to form condensed summaries that capture fine-grained differences among conditions of interest. Despite their inevitable limitations, then, they provide a standard of comparison against which other methods can be evaluated.

This paper pursues two aims. First, we provide an overview of how ordinal response variables are handled in the linguistic research literature. This survey looks at the kinds of ordered scales used and the way in which data are analyzed statistically. While MRMs are by far the most popular strategy, we find that ordered regression models are also used at a noticeable rate. The way in which their results are presented, however, fails to meet the standards set by MRMs – the typical report consists of a regression table with indications of statistical significance. Our survey suggests that language data analysts may not be aware of an alternative method of interpretation that is available for the type ordered regression model that permeates our research literature. This interpretative framework invokes an underlying continuous response variable and offers accessible model summaries that avoid the pitfalls tied to MRMs. The second goal of this paper is therefore to introduce linguists to this latent-variable formulation of ordered regression models. A tutorial for its implementation in R can be found in the online supplementary materials for this article.

With these two aims in mind, we have structured the remainder of this article as follows. In the next section, we summarize the results of our survey. Since MRMs are broadly used in the literature, Section 3 recapitulates the main limitations of this approach. In Section 4 we briefly introduce our illustrative data and clarify the analysis task. The data are then analyzed in two different ways. First, we use an MRM (Section 5) to address our research questions, and to emphasize its main assets. Section 6 then explains the rationale underlying ordered regression and its latent-variable formulation. This sets the scene for Section 7, which repeats the foregoing analyses using a latent-variable model (LVM). After a focused comparison of the results from both analyses (Section 8), Section 9 concludes with a summary.

2. Ordinal response variables in language data analysis

To start with, let us examine the treatment of ordered response variables in the linguistic literature. Following a description of our sample of research articles (Section 2.1), we give an overview of the scale layouts (Section 2.2) and analysis strategies (Section 2.3) that are used. In Section 2.4, we review relevant methodological work.

2.1. Selection of studies and overview of database

Our survey covers an 11-year span (2012–2022) and includes 3,859 articles published in 14 linguistic journals. These represent a broad range of research areas and methodologies (for an overview, please refer to Web appendix 1: https://osf.io/syavw). We searched for all studies that include the words "ordinal", "Likert", or "semantic differential", which returned 423 documents. Since our attention is restricted to the treatment of ordinal *outcome* (i.e. dependent or response) variables, we manually excluded work that features ordinal predictors only, or where ordinal variables play an ancillary role (e.g. for sample description or stimulus validation). We also excluded studies that rely on standardized (multi-item) tests to measure latent traits such as anxiety or motivation. This left us with 149 research articles.

The vast majority of ordinal responses in our database were collected using rating and judgment tasks (n = 136; 91%), where informants indicate some type of assessment or agreement by choosing from an ordered set of categories. Acceptability judgments were the most frequently employed scheme (n = 32). Recurrent forms also included accentedness and naturalness ratings, and semantic differential scales. For simplicity, we will refer to these kinds of response scales, collectively, as rating scales. Among the minor types of ordinal variables we encountered are existing categories (e.g. CEFR levels, phones/phonemes, or position on a grammaticalization cline) or classifications based on defined criteria (e.g. test or performance ratings).

2.2. Structure of ordinal scales

Let us take a closer look at the layout of response scales used for rating tasks (n = 136 articles). Table 1 summarizes the distribution of (i) number of response categories, ranging from 3 to 11; and (ii) the way in which descriptors are added to the scale, e.g. whether labels are given only at the endpoints, or for each category. The numbers in boldface provide the overall distribution of these attributes.

Considering the number of response categories, the most popular choices are 5 (39%) or 7 categories (35%); fewer than 5 and more than 7 options are rarely used. We also note that the vast majority of studies (79% in total) use an uneven number of tick boxes, thus providing a middle category that does not force ratings toward either end of the scale. As for the way in which descriptive information is added, we observe that 58% of all studies provide labels only at the endpoints. Another frequently used format is to give descriptions for each category (31%); around half of these fully labeled sequences are genuine Likert scales, where respondents indicate (degree of) agreement to a statement (Likert 1932).

Table 1. Structure of ordinal scales used in rating tasks (n = 136 articles).

	Number of response categories									
Labels	**3**	**4**	**5**	**6**	**7**	**9**	**10**	**11**	**Total**	
Endpoints only		3	24	8	29	2		2	**68**	**58%**
Each category	2	3	17	8	6				**36**	**31%**
Numbers			4	2			1		**7**	**6%**
Endpoints and midpoint			1		2			1	**4**	**3%**
Midpoint only							1		**1**	**1%**
Stars			1						**1**	**1%**
No information provided			6		10	3			(19)	
Total	**2**	**6**	**53**	**18**	**47**	**5**	**2**	**3**	**136**	
	1%	**4%**	**39%**	**13%**	**35%**	**4%**	**1%**	**2%**		

2.3. Analysis strategies used in the literature

Based on the analysis strategy used, we divided studies into six groups. An analysis is coded as employing a *mean response model* if it translates the ordered categories into numeric scores and then summarizes these using averages (and standard deviations), ANOVA or ordinary (mixed-effects) least-squares regression. The label *non-parametric test* was assigned to rank-based inferential procedures (e.g. Wilcoxon signed-rank test). Studies in the class *ordered regression model* employ a form of categorical regression respecting the order of response levels. Further, *ordered random forests* are listed as a separate category since they rely on a different modeling paradigm. Analyses assigned to the category *binary regression* dichotomize the graded scale, either by collapsing adjacent levels to form a binary outcome, or by applying a series of disconnected regressions to pairs (or groupings) of response levels. Finally, the label *description* is used for analyses that exclusively rely on descriptive statistics and use measures other than the mean (e.g. category counts/percentages or medians). We found 9 studies in our survey that rely on means (MRM) in addition to another strategy. These studies were assigned to both groups and appear twice in the following summaries, which are therefore based on 158 (instead of 149) data points.

Figure 1 shows the usage rate of strategies across the 11-year period covered by our survey. Each crossbar denotes a study, and the tallies are the total number of research articles making use of the respective approach. The great majority of analyses (n = 121; 77%) assign numeric scores to the ordered categories and then rely on MRMs for data summary and analysis. As for the choice of numeric scores, we found that, except for a single study, all analyses used equal distances between categories. The only other analysis strategy that is used at a notable rate is ordered regression (n = 23; 14%). All models reported in these papers are cumulative (mixed) models, a common variant of ordered regression that will be discussed in more detail in Section 6.

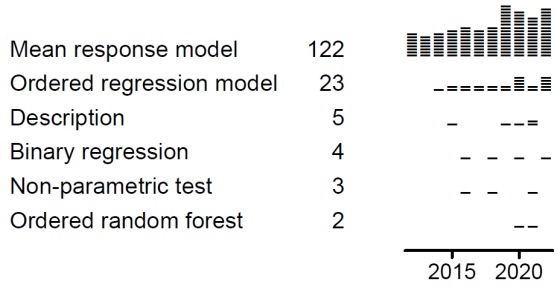

Figure 1. Strategies used to analyze ordinal outcome variables.

While ordered regression models featured in only 14% of the studies, this rate appears to be relatively high compared to other behavioral sciences (cf., e.g. Liddell and Kruschke 2018: 329). It is therefore of interest to also pay heed to the methodological literature directed at language scholars, which has succeeded in expanding our data-analytic repertoire.

2.4. The methodological literature on ordinal language data analysis

Ordered regression models are described in several of the statistical textbooks written for a broader linguistic audience. Thus, both Baayen (2008: 208–214) and Gries (2021: 353–361) illustrate how to run such models in R and how to summarize associations between predictor(s) and outcome using graphs of predicted category probabilities. A number of methodological articles have dealt with various aspects of ordinal data modeling: Endresen and Janda (2017) compare different approaches, including (mixed-effects) ordered regression modeling, ANOVA, and machine learning tools, i.e. regression and (unordered) classification trees with follow-up random-forest analyses. In their assessment, machine learning procedures evaluate most favorably in terms of model adequacy, informativity, and user-friendliness. Another paper by Baayen and Divjak (2017) illustrates how generalized additive mixed models can be applied to ordered outcomes. These capture non-linear associations between continuous predictor(s) and ordered response on an underlying latent variable scale (see Section 6.2). We note, in passing, that all methodological treatments mentioned so far concentrate on the same form of ordinal regression: the cumulative model (see Section 6).

An area of active methodological research is experimental syntax, which often deals with ordinal data derived from acceptability judgments (see Schütze and Sprouse 2014). This line of inquiry has directed attention to aspects such as the comparison of different measurement scales (e.g. binary choices, ordinal judgments, and magnitude estimation) and the agreement between formal and informal methods of data collection (e.g. Sprouse et al. 2013). Despite the ubiquity of ordinal scales in this field, the application of ordered regression has so far received little attention. Some recent work has explored the use of a latent-variable approach based on signal detection theory, both for binary judgments (Bader and Häussler 2010) and ordered scales (Dillon and Wagers 2021). For ordered responses, however, this analysis strategy can only yield difference measures on the underlying scale (Dillon and Wagers 2021: 89). Though informative for certain analysis tasks, this limits the wider applicability of the approach.

The current study concentrates on a more versatile latent-variable procedure for ordinal data analysis. Before we deal with this analysis strategy, however, let us recapitulate not only the limitations, but also the strengths of MRMs.

3. Limitations of mean response models

It has been pointed out in the statistical literature that the practice of converting an ordered response into numeric scores for analysis is problematic in certain settings. In this section, we summarize the main arguments that have been advanced against MRMs for ordered outcomes (see Long 1997: 35–40, 116–119; Agresti 2010: 5–8, 137–140).

The first major point of criticism is directed at the mapping between ordered categories and numeric values. While researchers are free in their choice of scores, equal-spaced integers appear to be a near-universal default (cf. Section 2.3). However, if verbal descriptions are given for categories (31% of the rating scales in our survey), perceived increments along the ordered scale will depend on how informants interpret these labels. Arguably, this concern is less of an issue with rating scales where only the endpoints are labeled (58%) and tick boxes are equally spaced, or where only numbers are used as category labels (6%).

A second major concern are floor and ceiling effects, which may arise with any form of ordinal (rating) scale. Thus, if responses, either in general or for certain subgroups, tend to be near the endpoints of the scale, the distribution of numeric scores will be skewed and measures of dispersion will be systematically smaller. Likewise, differences between conditions are compressed near the limits, which distorts data interpretation. This is especially true for interaction patterns, which may be scale-dependent, i.e. result from floor and ceiling effects (e.g. Loftus 1978; Dillon and Wagers 2021: 65–68).

A number of other problems have been noted. For instance, discrete scores make no allowance for the fact that responses are actually compatible with a range of values underlying the ordered scale. The discomfort caused by this form of measurement error grows if the number of response categories is small, which could be argued to apply to about half of the rating scales in our survey (with five or fewer levels). Further, MRMs may yield predictions (or uncertainty bounds) beyond the limits of the scale and they do not generate predicted category probabilities, which makes it difficult to check the fit of a model against the observed data.

Some limitations can be partly addressed at the design stage of a study. To mitigate floor and ceiling effects, for instance, the endpoints should mark extreme categories (e.g. "fully acceptable" instead of "acceptable"), and the number of categories can be increased (which also reduces measurement error). As for the choice of scores, sensitivity analyses can be used to check whether linguistic conclusions change with reasonable alterations of the numeric scores. To reinforce an equal-distance interpretation, Endresen and Janda (2017: 226–227) suggest adding numbers to tick boxes and to pay attention to the visual composition on the page/screen.

While these remedies may succeed in addressing some of the concerns we have reviewed in this section, they cannot guard against all potential problems associated with MRMs. Before we turn to ordered regression models as a more adequate analysis strategy, we introduce the linguistic context of our illustrative data.

4. Linguistic background and data

The data we rely on in this paper are questionnaire-based preference ratings for pairs of synonymous expressions (e.g. *package-parcel*). In Section 4.1 we describe the linguistic background and structure of our data. Section 4.2 outlines the research questions driving the analyses presented in the remainder of this article.

4.1. Background and data

Our data are drawn from a larger research project on lexical and grammatical variation in settings where English is spoken as a native, second, or foreign language. Here, we focus on the variety spoken and written in Malta and its relation to the major reference varieties of standard British and American English (BrE and AmE, for short). We concentrate on 68 pairs of lexical expressions such as *truck-lorry* or *realization-realisation*, which are known to differ in usage between BrE and AmE (cf. Algeo 2006). We will refer to these as lexical pairs, and to the individual expressions as variants. Though clearly a simplification, we will use BrE (or AmE, as the case may be) when we refer to 'more British', 'exclusively British' or 'traditionally British' items.

To elicit usage preferences, our questionnaire asks informants to indicate whether they always use one of the two variants, prefer one over the other, have no preference, or do not use either expression (see Appendix 1 for illustration; Krug and Sell 2013 for methodological details). We are therefore dealing with a symmetric 5-point ordinal scale with a meaningful midpoint and verbal descriptions for each response category. Data were collected between 2008 to 2018, and in this paper we concentrate on a subset of 500 speakers that is roughly balanced on year of birth (the terms

"speaker" and "informant" are used interchangeably in this paper). The distribution of this variable, which spans roughly 60 years, is shown in Figure 2, where each bar indicates the number of informants for a specific birth year. While the peak around the year 1990 results from the original data collection regime, it turns out that the age distribution for the Maltese population shows a similar amplitude for cohorts born around 1990.[1]

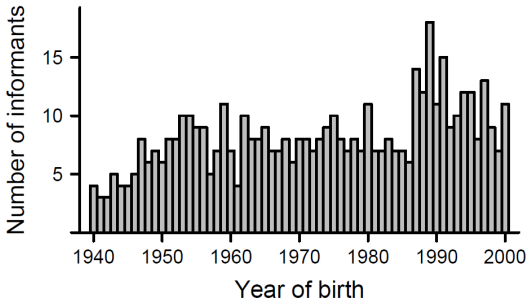

Figure 2. Distribution of year of birth in our sample of speakers.

Item pairs with more than 15% missing responses were excluded, leaving 63 pairs for analysis. The data[2] used in the present study are available from TROLLing (Krug et al. 2023). Appendix 2 shows, for each item, the distribution of the five response categories by year of birth. The color scheme adopted is used consistently in this article: Brighter shades denote the BrE variant, darker shades the AmE variant.

4.2. Research questions

The focus of our analysis is two-fold: First, we would like to assess, for each item pair, how much Maltese English tends toward the BrE or the AmE variant. When localizing a pair on the cline between the standard varieties, we will refer to its "Britishness", to acknowledge the historically more influential variety. Our second focus is on apparent-time trends for each item pair. This means that we are interested in the extent to which differences between age groups (as indicated by year of birth) may point to diachronic (i.e. real-time) changes in Maltese English. We would like to summarize apparent-time patterns using trend lines, which show us how usage preferences vary across cohorts. Both research questions can be addressed in a straightforward manner with an MRM. This approach will be illustrated in the next section.

5. Analysis using a mean response model

We first need to convert response categories to numeric scores. To facilitate interpretation, we assign the value 0 to "No preference", the middle category. In order for our numeric scale to reflect Britishness, we use positive scores to represent a preference for BrE and negative scores for AmE. Since the wording of the response categories (cf. Appendix 1) makes it difficult to reason about distances between these, we opt for equal-sized steps, i.e. scores of ±1 and ±2. The numeric scale is therefore bounded at −2 and +2, and zero assumes the interpretation of a draw between the standard varieties.

[1] https://www.populationpyramid.net/malta/2019/

[2] No ethics approval was obtained for this study, since – due to its design and participant characteristics – this was (and is, as of 2023) not required by European or Maltese national regulations or policies. For a careful weighing of research-ethical considerations, please refer to the data protection impact assessment in the TROLLing post (Krug et al. 2023).

For the analyses presented in this section, we use standard (mixed-effects) linear regression. To account for the data structure, all models include by-informant random intercepts, which allow for correlated responses within individuals. This provides leeway for variation in Britishness among speakers that is not captured by the predictors in a model. The fixed effects depend on the analysis task: To obtain overall Britishness scores, we include item pair as a predictor (model 1). Apparent-time trends, on the other hand, are captured with item pair, year of birth, and their interaction (model 2):[3]

- Model 1: `rating ~ item_pair + (1|informant)`
- Model 2: `rating ~ item_pair*year_of_birth + (1|informant)`

We used R (R Core Team 2023) and relied on the packages 'lme4' (Bates et al. 2015), 'emmeans' (Lenth 2023), and 'lattice' (Sarkar 2008). Details about the analysis can be found in the OSF project associated with this article (https://osf.io/jnv27).

Figure 3 shows the distribution of the 63 Britishness scores, which are estimated means[4] based on model 1. We observe that for the majority of lexical pairs, the speakers in our sample prefer the British variant – only in five cases do we note an appreciable preference for the AmE expression. There is considerable variation among item pairs in terms of the strength of preference. The distribution hits the upper bound of the scale, with some item pairs near ceiling, indicating almost categorically British usage.

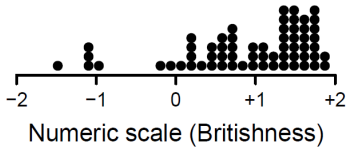

Figure 3. Gradience in Britishness: Estimated means for the 63 item pairs, based on an MRM (model 1).

MRMs also allow us to summarize apparent-time trends in the data. Figure 4 illustrates this for *package-parcel*, which shows the clearest trend away from BrE. In this graph, year of birth appears on the *x*-axis, and each open circle marks the rating given by a specific informant. The vertical location of points is pulled apart somewhat to avoid overlap, which allows us to see more clearly where ratings gravitate across cohorts: While informants born around 1950 tend to prefer the British variant, this pattern changes across age groups, with speakers born in the 1990s leaning toward the AmE form. The line in the graph shows a straight-line summary of the estimated trend.[5]

[3] In the notation used here, the term in brackets denotes the random intercepts for the variable informant.
[4] These are the fitted values obtained through appropriate combination of the regression coefficients (i.e. the fixed intercept and slopes).
[5] This is the regression line fitted to the (unjittered) data.

Figure 4. Data representation for item pair *package-parcel* using an MRM; trend line showing estimated means by year of birth (model 2).

The ability to summarize trends using simple profiles allows us to take a bird's eye perspective and bring into view all item pairs in our data. To this end, Figure 5 collects apparent-time patterns (i.e. fitted regression lines) from model 2 in a single display. It appears that most pairs show a downward trend – similar to *parcel-package*, though not as pronounced. Note that the ceiling effect, which was evident in Figure 3 above, also surfaces in this graph – most trend lines hover at the British end of the scale.

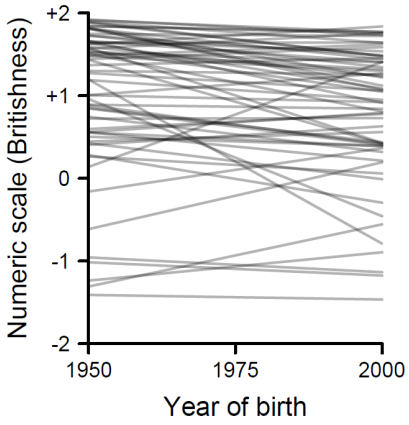

Figure 5. Apparent-time trends for the 63 item pairs based on an MRM (model 2)

To simplify our assessment of the patterns shown in Figure 5 yet further, we can summarize each trend line by taking the difference (or contrast) between the estimated means for speakers born in 1950 and speakers born in 2000.[6] Negative differences then denote a decrease in Britishness, with the magnitude reflecting the steepness of the trend lines. Figure 6, which graphs these contrasts in rank order, shows that most are negative. Our illustrative item pair (*package-parcel*) appears at the far left, with an estimated apparent-time decrease of 2 points (cf. Figure 4). Item IDs (running from 1 to 63) have been added to the display, and reference to Appendix 2 allows us to look up pairs of

[6] In the current linear regression model, these contrasts of estimated means correspond to the item-specific slopes for year of birth.

expressions. To assess the statistical dependability of these apparent-time differences, we have added 95% confidence intervals.

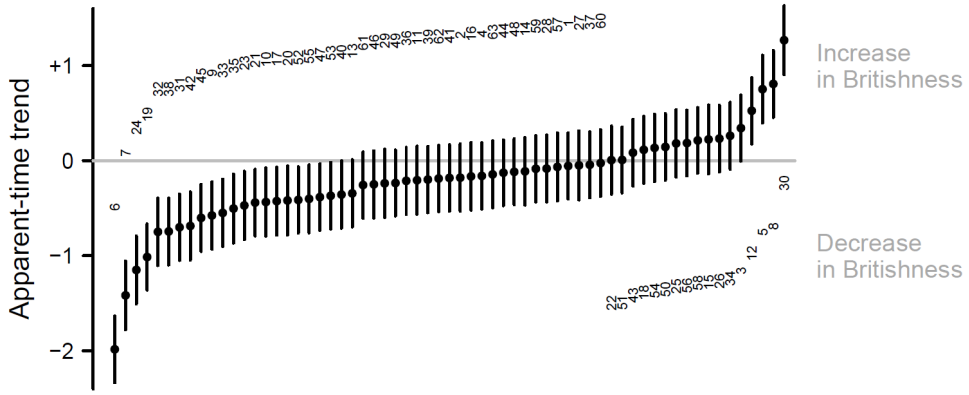

Figure 6. Apparent-time trends based on an MRM: Contrasts of estimated means for birth years 2000 and 1950 with 95% confidence intervals (model 2).

The analyses in this section demonstrate how MRMs allow us to form simple and informative data summaries: The means and trend lines we have presented provide direct answers to our research questions. In terms of model interpretation, then, MRMs provide a benchmark against which alternative analysis strategies may be compared.

6. Ordered regression models

Three basic types of ordinal regression are usually distinguished, based on the way in which (changes in) category probabilities are modeled: the cumulative, the adjacent-category, and the continuation-ratio model (see Bürkner and Vuorre 2019 for an overview; Fullerton and Xu 2017 for a book-length treatment). We will concentrate on the most commonly used version, the cumulative model (McKelvey and Zavonia 1975; McCullagh 1980). Similar to other categorical regression models, cumulative models do not describe response probabilities on the data scale (probabilities), but on an unconstrained model scale, which has no upper or lower limit. Two common choices are logit and probit scores.

It turns out that the cumulative model can be understood (or derived) in two distinct ways: as a representation of (differences in) model-scaled cumulative response probabilities, and as a description of a latent variable underlying the ordinal scale. Our focus in this paper is on the continuous response formulation (Section 6.1), and Sections 6.2 and 6.3 illustrate how analogues of the means, patterns and differences presented in Section 5 can be constructed on the latent-variable scale. All ordinal models reported in this paper were run in R using the package ‘ordinal’ (Christensen 2022).

6.1. Latent-variable representation of cumulative models

The latent data formulation of cumulative models stipulates a continuous variable underlying the sequence of ordered responses (cf. Long 1997: 116–122; Agresti 2010: 53–55). We will stick to the label Britishness to designate this latent variable, and we will scale it in a way that positive scores reflect an inclination toward the British variant, with zero assuming the same interpretation as in the MRM.

To illustrate how a latent-variable model (LVM) negotiates between the ordered response categories and the continuous response formulation, let us again consider a hypothetical item pair. Let us assume we obtain, as an estimate of the propensity toward either standard variety, an average of +0.3. This average leans toward BrE, and Figure 7a shows that it marks the center of a bell-shaped distribution. This curve represents the variation in propensities underlying the responses we have collected using our questionnaire. This variability is the sum of different sources of variation. Possibly the greatest share is variation among respondents, as speakers may differ in their (strength of) preference – either in general, or for this lexical pair specifically. Other sources of variation that contribute to the spread of scores is error, perhaps due to fatigue or a misunderstanding of the meaning of the expressions.

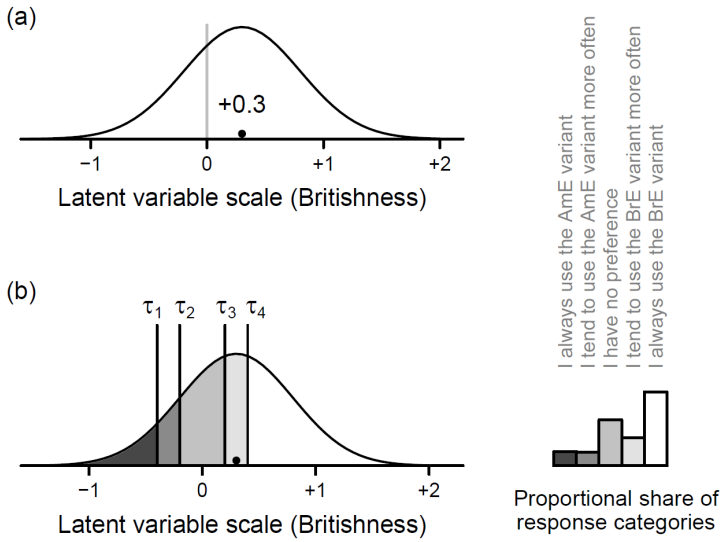

Figure 7. Illustration of the latent-variable formulation of the cumulative model.

The way in which this underlying latent scale maps onto the ordered response categories is illustrated in Figure 7b. The vertical lines mark four thresholds, which are labeled using the Greek letter τ (tau). These divide the continuum into five intervals, one for each category. For instance, if the latent Britishness score underlying a specific response in our questionnaire (and recall that this score may include an error, or noise component) is +1.0, the response category "I always use the British term" will be selected.

The location of the thresholds is determined empirically, by the regression analysis. Their location and spacing depends on the number of response categories and on how they are labeled, both in absolute terms and relative to each other. The estimation of threshold parameters is usually flexible, which means that they are allowed to adapt to the distribution of the data, without any a-priori constraints (e.g. equidistance). For our 5-point scale, we implement two constraints: First, since the middle category is neutral, we center thresholds at 0. The sign of the latent variable is then meaningful, since positive averages signal a preference for BrE. Further, with our response categories being identically phrased at both ends, we constrain thresholds to be symmetric around 0. Irrespective of how thresholds are estimated, they divide the area under the density curve into five parts, which are represented in Figure 7b with different fill colors. These areas under the curve represent the model-based estimates of the category probabilities. They are shown to the right of the graph using a more familiar graph type, a grouped bar chart.

Depending on the specific distribution that is used to represent the underlying variability, we distinguish (among others) between the cumulative model with a *probit* link (normal distribution)

and a *logit* link (logistic distribution). The choice between these two is largely a matter of convenience (Long 1997: 120); we will use the probit link function. Next, we demonstrate how the latent-variable approach allows us to construct model summaries that meet the standards set by MRMs.

6.2. Comparison of items on the latent variable scale

Figure 8 illustrates how the responses for three item pairs in our data are modeled, with the *x*-axis hosting the latent variable, Britishness. The normal densities have different means: For the item pair (a) *trash-rubbish* it is +1.2; for (b) *center-centre* it is +0.5; and for (c) *package-parcel* it is +0.2. Four thresholds divide the horizontal scale into five regions, which represent our response categories. The grouped bar charts at the right show the estimated response probabilities. Note how, for *package-parcel*, the preference pattern is roughly balanced: The share of the outer categories ("I always use this expression") are 29% for *package* (AmE) and 40% for *parcel* (BrE). For *center-centre*, the latent variable mean is higher, and we can observe how the proportional share of blue and red change accordingly. The proportion of speakers "always" using BrE *centre* is 60%; for BrE *rubbish*, it is at 79%. This illustrates the effect of the latent-variable mean on the category probabilities: As it shifts along the scale, we observe a change in response probabilities.

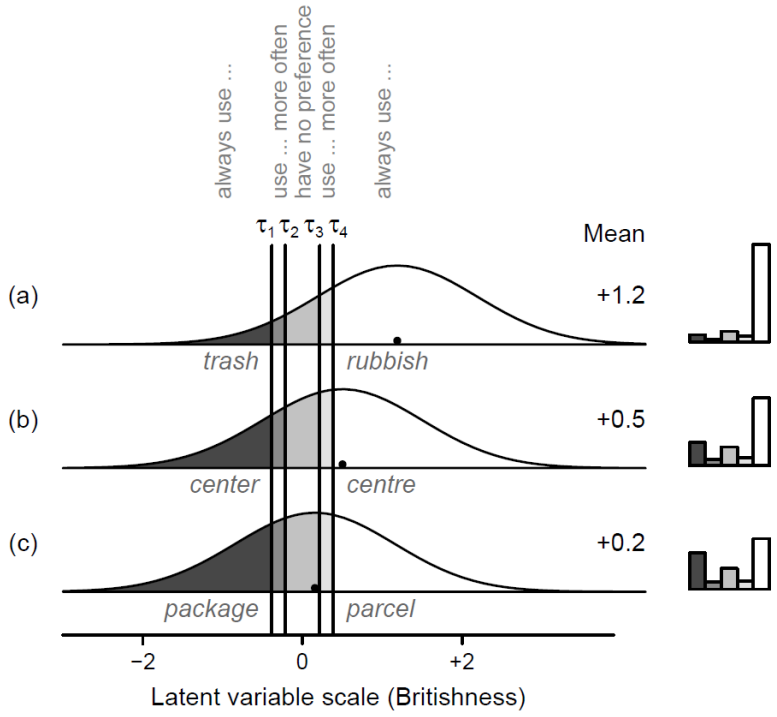

Figure 8. LVM for three illustrative item pairs in our data set.

Let us also consider how this information is represented in a regression table. For the kind of analysis underlying Figure 8, there are three types of parameters: (i) thresholds, (ii) coefficients describing the item means, and (iii) a random intercept SD, which expresses between-speaker variability in the overall level of Britishness. Further, Table 2 shows two sets of thresholds: An uncentered and a centered one (with mean 0). The latter parameters are the ones underlying Figure 8. The difference between these sets of thresholds is a shift in scale: In the centered set, the average over the uncentered parameters (−0.67) is subtracted out.

Table 2. Table of coefficients for the analysis underlying Figure 8.

Coefficient	Estimate	(SE)	Centered†
Thresholds			
τ_1 (1|2)	−1.00	(0.05)	−0.39
τ_2 (2|3)	−0.83	(0.05)	−0.22
τ_3 (3|4)	−0.40	(0.04)	+0.22
τ_4 (4|5)	−0.23	(0.04)	+0.39
Coefficients			
Item pair (sum contrasts)			
Contrast 1: (a) – grand mean	−0.46	(0.05)	
Contrast 2: (b) – grand mean	−0.11	(0.05)	
Random intercept SD (Informant)	0.40		

Note. † These thresholds are centered around zero by subtracting out −0.61, the midpoint between thresholds 2 and 3. Against these centered thresholds, the predicted latent-variable means are therefore shifted upward by +0.61.

6.3. Apparent-time trends on the latent variable scale

An LVM can also be used to describe trends in the data. We are then interested in how the estimated latent-variable mean for a specific item pair varies with year of birth. A graphical representation of such a model for the item pair *package-parcel* appears in Figure 9. Britishness is now shown on the vertical axis, which is again divided into five intervals by the thresholds. The normal distributions in the graph represent six equally-spaced snapshots, starting with the year 1950 (left) and extending to the year 2000 (right). The tilted line connects the estimated means at these x-values, and (in accordance with the model assumptions) it forms a straight line. Estimated means are given above the graph – they range from +0.9 (1950) to −0.6 (2000). The grouped bar charts at the top show the predicted category probabilities for different birth years: The share of responses that indicate "always" using *parcel* decreases from 67% (1950) to 16% (2000). Conversely, the proportion of "always"-responses for *package* increases from 10% to 57%.

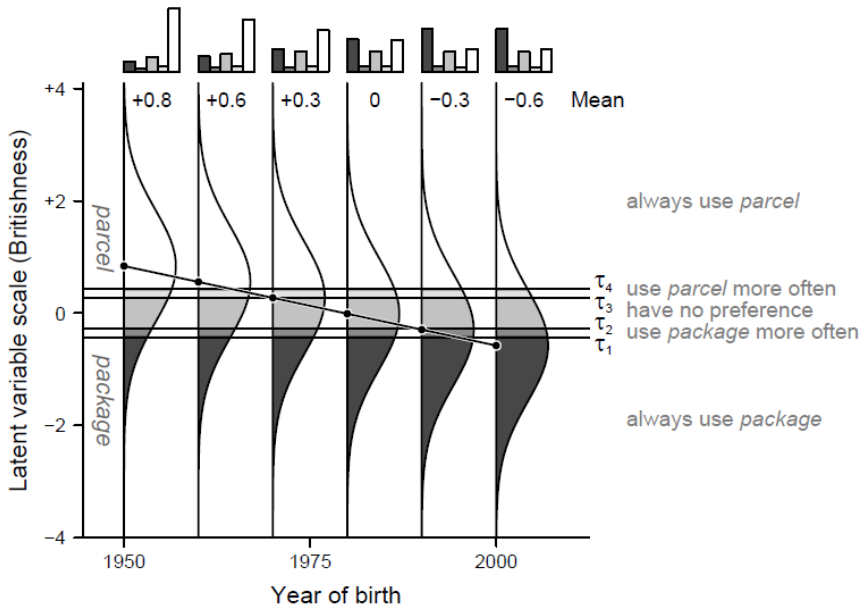

Figure 9. LVM for the apparent-time trend in the item pair *parcel-package*.

Let us also consider a regression table for the analysis shown in Figure 9, where we now have two types of parameters: thresholds and a coefficient describing the trend for *package-parcel*. Again, both uncentered and centered thresholds are shown.

Table 3. Table of coefficients for the analysis underlying Figure 9.

Coefficient	Estimate	(SE)	Centered[†]
Thresholds			
τ_1 (1|2)	−0.56	(0.06)	−0.43
τ_2 (2|3)	−0.40	(0.06)	−0.27
τ_3 (3|4)	+0.14	(0.06)	+0.27
τ_4 (4|5)	+0.30	(0.06)	+0.43
Coefficients			
Year of birth, centered[‡]	−0.71	(0.08)	

Notes. [†] These thresholds are centered at zero by subtracting out −0.13, the midpoint between τ_2 and τ_3. The estimated means in Figure 9 are therefore shifted upward by +0.13.
[‡]Centering: (Year − 1975)/25; −1 corresponds to 1950, 0 to 1975, and +1 to 2000

The model-based change in category probabilities can also be shown on a continuous scale using an area chart. This graphical arrangement appears in Figure 10a, where we see the continuous decrease in Britishness in apparent time. For comparison, the descriptive density plot in panel b offers a smoothed summary of category probabilities across birth years (cf. Appendix 2). We see that the fit of the model is not perfect – most notably, the share of "no preference" responses appears to be underestimated somewhat.

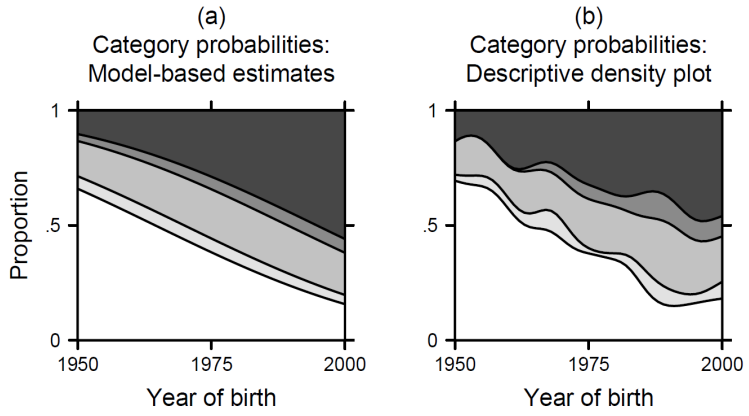

Figure 10. Visualization of the apparent-time trend in *parcel-package*: (a) Model-based category probabilities; (b) a flexible summary using a conditional density plot.

We have seen how differences between items and apparent-time patterns can be modeled and summarized on the latent-variable scale. In the next section, we apply these techniques to the full set of data.

7. Analysis using a latent-variable model

We now use a cumulative mixed model with a probit link to address our research questions. We will first return to the gradience in Britishness among item pairs, and then consider apparent-time trends in our data. For comparison, we will present results side-by-side with those obtained from the earlier MRMs. In terms of structure (fixed and random components), the ordered regression models we report here parallel those in Section 5; this means that we used the same model syntax (models 1 and 2).

Figure 11b shows the latent variable means for the 63 items. The thresholds, which guide the interpretation of these scores, also appear in the graph. Similar to the MRM-based distribution (panel a), we note that the majority of items sits firmly in the British half of the continuum. Since the latent scale is unconstrained, however, the pile of dots shows no ceiling effect and instead forms a more symmetric distribution.

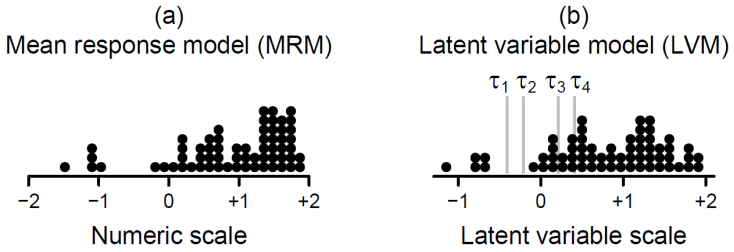

Figure 11. Gradience in Britishness: Variation among the 63 item pairs, based on (a) the MRM and (b) the LVM (model 1).

Figure 12b displays the set of 63 straight-line patterns based on the LVM, along with the four thresholds. A comparison to panel (a) reveals two differences: First, the trend lines are spread out more evenly along the scale, which improves resolution among item pairs with high levels of Britishness. Further, the removal of the ceiling effect sensitizes the model to differences in trend at the upper end of the scale. While in panel (a) there is little variability in slopes among items near ceiling, no such scale constraints are evident in panel (b).

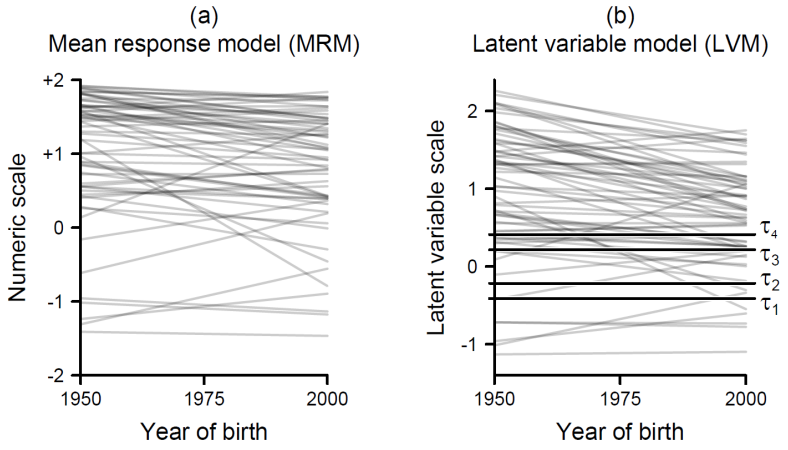

Figure 12. Apparent-time trends: Variation among the 63 item pairs, based on (a) the MRM and (b) the LVM (model 2).

Let us also look at a condensed summary of the 50-year apparent-time trends. Figure 13 shows contrasts of estimated means, i.e. differences in Britishness between 2000 and 1950. Recall that positive values indicate an increase in Britishness, and negative scores reflect a decrease. A comparison to Figure 6 shows a somewhat greater number of confidence intervals that do not cover zero. We will take a closer look at these differences in the next section.

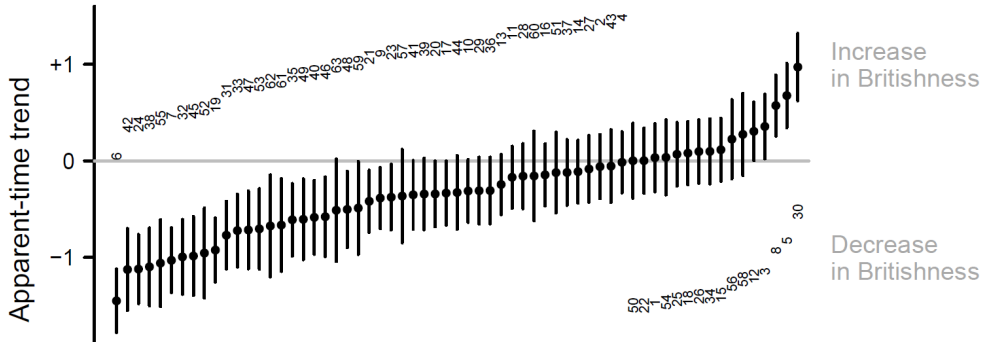

Figure 13. Statistical uncertainty estimates for apparent-time trends based on an LVM (model 2). Error bars denote 95% confidence intervals.

8. Comparison of results

We now present more detailed, item-level comparisons of the results from the two models, dealing in turn with the overall level of Britishness and apparent-time trends.

Figure 14 shows estimated means – or Britishness scores – from the two models. Each item pair is represented by a line, which connects the estimated MRM (top) and LVM (bottom) mean. The two scales are aligned at zero, which has the same interpretation in both models. Apart from this alignment, we have scaled the horizontal extension of each set of scores based on their standard deviation. This makes them roughly comparable in terms of visual spread. The axes, however, show the actual scores (i.e. not the standardized[7] scores).

What emerges from Figure 14 is that the Britishness scores from the two models are in very good agreement (Pearson correlation: +.99). The rank order of item pairs is almost identical, which is evident from the fact that only few profiles cross. Figure 14 also illustrates how the latent scale increases the resolution among item pairs with high Britishness scores; it is more successful in revealing gradience near the extremes of the scale.

[7] These scores are in fact only partly standardized, since they are not centered about their mean.

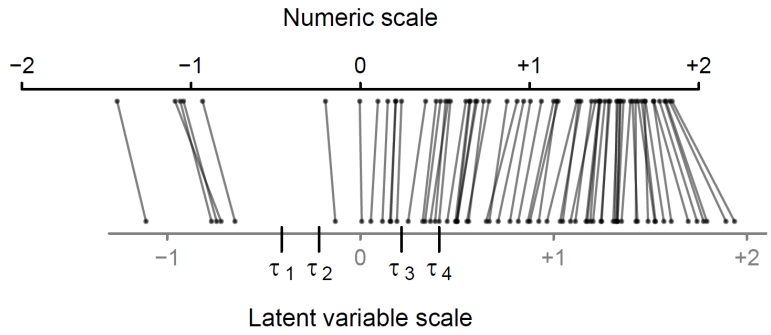

Figure 14. Comparison of estimated means (Britishness scores) of all 63 item pairs based on an MRM (top) and an LVM (bottom) (model 1).

Figure 15 compares the estimates for apparent-time differences – i.e. contrasts of estimated means – derived from the two models. The setup is the same as in Figure 14: The scales are aligned at zero (indicating no change in apparent time) and the horizontal spread of estimates is standardized, with the axis labels giving the actual differences. With a Pearson correlation of +.90, the overall agreement between the two models is still good. In comparison to Figure 14, however, we note more crossings of profiles. As we draw lines from the top (MRM) to the bottom (LVM), a recurrent pattern is for negative contrasts (signaling an apparent-time decrease in Britishness) to show a leftward deflection. In other words, for a number of items the LVM yields steeper trend lines away from the BrE variant.

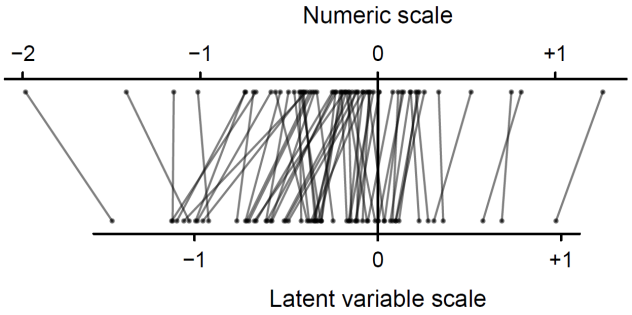

Figure 15. Comparison of contrasts of estimated means (our estimates of apparent-time trends) for all 63 items, based on an MRM (top) and an LVM (bottom) (model 2).

Let us take another look to understand this pattern. To this end, Figure 16 shows, on the left-hand side, the differences between the standardized apparent-time contrasts of estimated means derived from the two models. A dot below the grey horizontal line indicates that, for this specific item pair, switching from the MRM to the LVM brings about a clockwise rotation of the (standardized) trend line. Since the majority of these (second) differences are negative, this is the typical rotation we observe. At the right-hand side, Figure 16 graphs these differences in apparent-time trend against the estimated latent-variable mean in Britishness (using item IDs from Appendix 2 as plotting symbols). It provides an exploded view of the dot diagram and allows us to see, for each item, how the observed between-model difference in apparent-time trend relates to the overall level of (latent-scale) Britishness. We note that a clockwise rotation of trend lines tends to go hand in hand with a greater average level of Britishness. In other words, items with a high level of Britishness show a steeper decrease in Britishness in the LVM. This results from a removal of the ceiling effect: While MRM trend lines are artificially flattened near the end of the scale (see Figure 12a), the LVM removes this constraint and allows our straight-line summaries to stretch out, providing a better picture of the behavior of firmly British item pairs.

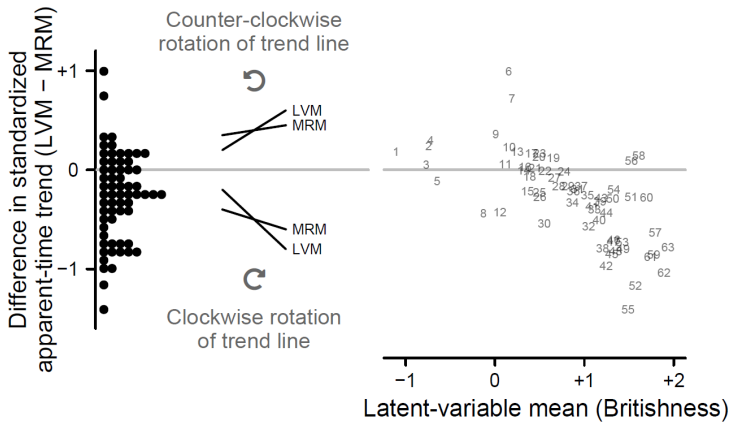

Figure 16. Discrepancy between standardized contrasts of estimated means (standardized apparent-time trends) generated by the two models, and its relation to the endpoints of the scale, i.e. the estimated latent-scale mean (Britishness score).

Let us finally compare the statistical uncertainties in the trend coefficients. Figure 17 shows the contrasts of estimated means produced by the two models, along with 95% confidence intervals. Estimates from the LVM (cf. Figure 13) are shown in grey, and the item pairs are ordered based on these. The vertical position of IDs reflect the estimated latent-scale means of Britishness; most therefore appear in the upper half of the display. The squares at the top of the graph flag items whose 95% interval does not cross zero, i.e. items that might tentatively be considered as showing a statistically detectable trend; again, grey squares refer to the LVM.

For the MRM (black squares), 26 uncertainty intervals exclude zero, and for the LVM 32 item pairs show a detectable trend in apparent time. For the majority of item pairs, then, we arrive at the same binary statistical conclusion if such were needed. For 8 item pairs, only one of the intervals excludes zero. If we compare these mismatches to the latent-scale Britishness of items (vertical position of ID), we note two things: First, the LVM appears to be statistically more sensitive near the endpoints of the scale (i.e. solitary grey squares go hand in hand with high latent-scale means, i.e. item IDs appearing near the top of the graph). Further, the MRM more readily detects non-zero patterns near the scale midpoints (i.e. the solitary black square, for item pair 10, is linked to a latent-scale mean near zero).[8] Overall, however, Figure 17 shows near-perfect agreement between the models when it comes to the directionality of apparent-time patterns. There are no item pairs for which the two analyses yield qualitatively different conclusions.

[8] This is due to the fact that the MRM relies on a combined (i.e. pooled) estimate of residual variation around the conditional averages. Since the variation of scores is smaller near the endpoints of the scale (cf. Section 3), the residual standard deviation is downwardly biased for conditional means near the scale midpoint. The opposite is true, of course, for estimates near the bounds of the scale.

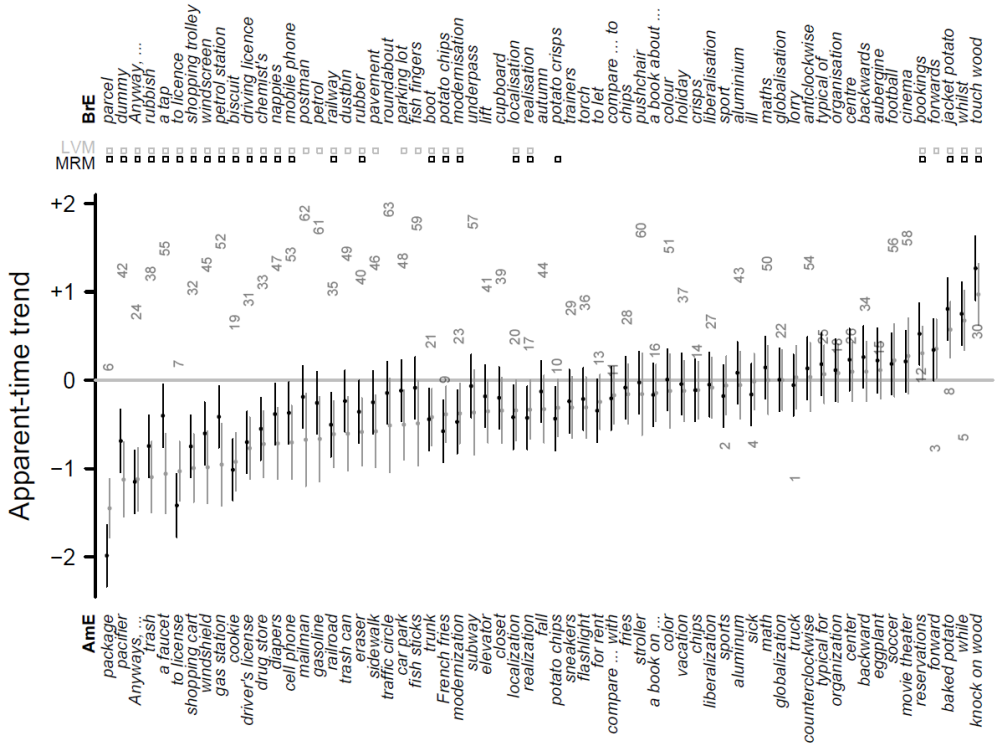

Figure 17. Statistical uncertainty estimates for apparent-time trends, based on an MRM (black) and an LVM (grey).

9. Summary and conclusion

With the statistical analysis of ordinal variables being a recurrent theme in the methodological literature, we conducted a survey of research articles published in various linguistic journals to see how they are handled in language data analysis. We observed that ordered responses are most commonly treated in the same way as continuous variables. While limitations of this approach have been pointed out in the literature, it is perhaps equally important to recognize its strengths: Apart from being easy to implement, mean response models (MRMs) provide informative and accessible data summaries. Our survey also showed that ordered regression models are making inroads into the published literature, with cumulative models (a specific form of ordinal regression) being applied in an increasing share of analyses. The way in which such models are summarized and interpreted in these studies suggests that not many linguists may be familiar with their alternative, latent-variable formulation. Critically, this interpretative angle allows us to summarize data patterns analogously to MRMs.

The aim of this paper was to draw linguists' attention to this conceptualization of cumulative models, and to illustrate its application to language data. We used data from a research project on lexical variation in Maltese English to juxtapose analyses based on MRMs and latent-variable models (LVMs). With regard to clarity of interpretation, the data summaries provided by the two model types were at eye level. At the same time, however, LVMs allowed us to avoid the limitations inherent in MRMs. Thus, they sidestep the difficult (and necessarily arbitrary) choice of numeric scores for the response categories, they cushion certain forms of measurement error, and they allow us to generate predicted response probabilities, which make it possible to compare model estimates to the observed data. LVMs also succeeded in removing from our model summaries distortions induced by floor and ceiling effects. Thus, for MRMs we observed how variation among simple averages and regression lines was compressed near the endpoints of the scale, with trend lines being

artificially flattened. While for overall Britishness scores this only led to some deformity in the relative spacing along the scale, the amount of change in apparent time suggested by our analysis varied systematically between the two models. The scale-dependence of interaction patterns therefore affected a quantity of immediate linguistic interest.

We hope to have illustrated how the latent-variable conceptualization of cumulative models aids interpretability and allows us to communicate ordered regression models in a manner that is both informative and accessible. Judging from the results of our survey, which indicates that the vast majority of studies either use a mean response model or a cumulative ordered regression model, we believe that an LVM approach to ordered response variables may be of wider applicability in language data analysis.

References

Agresti, Alan. 2010. *Analysis of ordinal categorical data*. Hoboken, NJ: John Wiley & Sons.

Algeo, John. 2006. *British or American English: A handbook of word and grammar patterns*. Cambridge: Cambridge University Press.

Baayen, R. Harald. 2008. *Analyzing linguistic data: A practical introduction to statistics using R*. Cambridge: Cambridge University Press.

Baayen R. Harald & Dagmar Divjak. 2017. Ordinal GAMMs: A new window on human ratings. In Makarova, Anastasia, Stephen M. Dickey & Dagmar Divjak (eds.), *Each venture a new beginning: Studies in honour of Laura A. Janda*, 39–56. Bloomington: Slavica Publishers.

Bader, Markus & Jana Häussler. 2010. Toward a model of grammaticality judgments. *Journal of Linguistics* 46(2). 273–330.

Bates, Douglas, Martin Maechler, Ben Bolker & Steve Walker. 2015. Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software* 67(1). 1-48. doi:10.18637/jss.v067.i01.

Bürkner, Paul-Christian & Matti Vuorre. 2019. Ordinal regression models in psychology: A tutorial. *Advances in Methods and Practices in Psychological Science* 2(1). 77–101. https://doi.org/10.1177/2515245918823199

Christensen, Rune H. B. 2022. ordinal - Regression Models for Ordinal Data. R package version 2022.11-16. https://CRAN.R-project.org/package=ordinal.

Dillon, Brian & Matthew W. Wagers. 2021. Approaching gradience in acceptability with the tools of Signal Detection Theory. In Grant Goodall (ed.), *The Cambridge handbook of experimental syntax*, 62–96. Cambridge: Cambridge University Press.

Fullerton, Andrew S. & Jun Xu. 2017. *Ordered regression models: Parallel, partial, and non-parallel alternatives*. Boca Raton, FL: CRC Press.

Gries, Stefan Th. 2021. *Statistics for linguistics with R*. Berlin: Mouton de Gruyter.

Janda, Laura A. & Anna Endresen. 2017. Five statistical models for Likert-type experimental data on acceptability judgments. *Journal of Research Design and Statistics in Linguistics and Communication Science* 3(2). 217–250. https://doi.org/10.1558/jrds.30822

Krug, Manfred & Katrin Sell. 2013. Designing and conducting interviews and questionnaires. In Manfred Krug & Julia Schlüter (eds.), *Research methods in language variation and change*, 69–98. Cambridge: Cambridge University Press.

Krug, Manfred, Ole Schützler & Valentin Werner. 2016. Patterns of linguistic globalization: Integrating typological profiles and questionnaire data. In Olga Timofeeva, Anne-Christine

Gardner, Alpo Honkapohja & Sarah Chevalier (eds.), *New approaches to English linguistics: Building bridges*, 35–66. Amsterdam: Benjamins.

Krug, Manfred, Fabian Vetter & Lukas Sönning. 2023. *Dataset for "Latent-variable modeling of ordinal outcomes in language data analysis"*. https://doi.org/10.18710/WI9TEH, DataverseNO, V1. An anonymized version of the dataset is available at https://dataverse.no/privateurl.xhtml?token=69cf2538-1127-4aa1-970f-1b8fa5e3e249.

Lenth, Russell V. 2023. emmeans: Estimated marginal means, aka least-squares means. R package version 1.8.5, https://CRAN.R-project.org/package=emmeans

Liddell, Torrin M. & John K. Kruschke. 2018. Analyzing ordinal data with metric models: What could possibly go wrong? *Journal of Experimental Social Psychology* 79. 328–348.

Likert, Rensis. 1932. *A technique for the measurement of attitudes*. New York: Columbia University Press.

Loftus, Geoffrey R. 1978. On interpretation of interactions. *Memory & Cognition* 6(3). 312–319. https://doi.org/10.3758/BF03197461

Long, J. Scott. 1997. *Regression models for categorical and limited dependent variables*. Thousand Oakes, CA: Sage.

McCullagh, Peter. 1980. Regression models for ordinal data. *Journal of the Royal Statistical Society Series B* 42(2). 109–142.

McKelvey, Richard & William Zavoina. 1975. A statistical model for the analysis of ordinal dependent variables. *Journal of Mathematical Sociology* 4. 103–120.

R Core Team. 2023. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Sarkar, Deepayan. 2008. *Lattice: Multivariate data visualization with R*. New York: Springer.

Schütze, Carson T. & Jon Sprouse. 2014. Judgment data. In Robert J. Podesva & Devyani Sharma (eds.), *Research Methods in Linguistics*, 27–50. Cambridge: Cambridge University Press.

Sprouse, Jon, Carson T. Schütze & Diogo Almeida. 2013. A comparison of informal and formal acceptability judgments using a random sample from *Linguistic Inquiry* 2001–2010. *Lingua* 134. 219–248.

Stevens, Stanley S. 1946. On the theory of scales of measurement. *Science* 103(2684). 677–680. https://doi.org/10.1126/science.103.2684.677.

Appendix 1. Excerpt from questionnaire

	I always use this expresion	I use this expression more often	I have no preference	I use this expression more often	I always use this expression		I never use either expression	Explanation/Comment
to licence	○	○	○	○	○	to license	○	
elevator	○	○	○	○	○	life	○	
localisation	○	○	○	○	○	localization	○	
truck	○	○	○	○	○	lorry	○	*(large motor vehicle for carrying goods by road)*
maths	○	○	○	○	○	math	○	
cell phone	○	○	○	○	○	mobile phone	○	
modernisation	○	○	○	○	○	modernization	○	
diapers	○	○	○	○	○	nappies	○	*(for babies)*
organisation	○	○	○	○	○	organization	○	
package	○	○	○	○	○	parcel	○	*(something you send by mail)*
pavement	○	○	○	○	○	sidewalk	○	*(for pedestrians, next to street)*
gasoline	○	○	○	○	○	petrol	○	

Appendix 2. Data distribution for item pairs. These area charts show smoothed conditional densities produced with the R function cdplot(), setting the bandwidth selection argument to "SJ". Tick marks along the x-axis mark the birth years 1950 and 1975.