This is an Accepted Manuscript of an article accepted for publication in *Research Methods in Applied Linguistics* (Elsevier).

Ordinal response scales: Psychometric grounding for design and analysis

Lukas Sönning (University of Bamberg)

Abstract. Ordinal response scales are commonly used in applied linguistics. To summarize the distribution of ratings or judgments provided by informants, these are usually converted into numbers and then averaged or analyzed with ordinary regression models. This approach has been criticized in the literature; one caveat (among others) is the assumption that distances between categories are known. The present paper illustrates how empirical insights into the perception of response labels may inform the design and analysis stage of a study. We start with a review of how ordinal scales are used in linguistic research. Our survey offers insights into typical scale layouts and analysis strategies, and it allows us to identify three commonly used rating dimensions (agreement, intensity, and frequency). We take stock of the experimental literature on the perception of relevant scale point labels and then demonstrate how psychometric insights may direct scale design and data analysis. This includes a careful consideration of measurement-theoretic and statistical issues surrounding the numeric-conversion approach to ordinal data. We focus on the consequences of these drawbacks for the interpretation of empirical findings, which will enable researchers to make informed decisions and avoid drawing false conclusions from their data. We present a case study on *yous(e)* in two varieties of English, which shows that reliance on psychometric scale values can alter statistical conclusions, while also giving due consideration to the key limitations of the numeric-conversion approach to ordinal data analysis.

Key words. Ordinal data, rating scales, acceptability judgments, Likert scale, judgment task, measurement, psychological scaling

1. Introduction

Ordinal response scales are commonly used across different branches of linguistics to elicit some kind of judgment, perception, or opinion. In experimental syntax, for instance, participants may rate the acceptability of a sentence on a 7-point scale ranging from "not at all acceptable" to "fully acceptable". When analyzing data obtained in this way, researchers routinely assign numbers to the categories (in this case, say, running from 1 to 7). Following Labovitz (1967), we will use the term *scoring system* to refer to the numeric translation of an ordinal scale. The data are then usually treated as though they had been measured on a continuous (or interval) scale, to calculate averages or use ordinary (mixed-effects) regression. This practice, which is widespread across empirical disciplines (see, e.g., Harwell & Gatti, 2001; Liddell & Kruschke, 2018) including linguistics (see Sönning et al., 2024), has sparked heated methodological debates (see, e.g., Harpe, 2015; Jamieson, 2004; Knapp, 1990; Norman, 2010; for linguistic data, see Endresen & Janda, 2017). The widely accepted belief that an interval-level analysis of ordinal data is inappropriate goes back to an influential paper by Stevens (1946), who proposed a taxonomy of scale types (nominal, ordinal, interval, and ratio) along with "appropriate" statistics for each. Among the caveats of the numeric-conversion approach is the fact that it requires information on the spacing between consecutive categories. Even though these distances are usually unknown, the statistical analysis proceeds as though they had been measured.

A survey of the use of ordinal response scales in language research shows that two broad types may be distinguished (see Section 2; also Krosnick & Fabrigar, 1997, p. 149): (i) fully verbalized sequences with labels for all categories; and (ii) scales that consist of a number of boxes with descriptors only at the endpoints. If only the extremes are labeled and the intermediate options are evenly spaced (on the page or screen), the equidistance assumption may be tenable. When all scale points are

verbalized, however, the perceived distance between categories will mainly depend on how informants interpret the labels; to encourage an equally-apportioned interpretation, however, numbers may be add to the scale point labels. To the advantage of empirical researchers, experimental studies have produced insights into the perception of quantificational expressions linked to a number of dimensions that are frequently used to build graded scales. Psycholinguistic research on intensifiers, for instance, has shown that English native speakers recognize similar increments in intensity between *hardly-slightly* and *considerably-highly* (Rohrmann, 2007). The aim of the present paper is to illustrate how such insights into the perception of response labels may inform study design and data analysis. We therefore build on earlier methodological work, which has mainly focused on scale construction (i.e., the selection of approximately equal-interval sequences; e.g., Beckstead, 2014; Friedman & Amoo, 1999; Rohrmann, 2007). The present study goes further, however, and considers how psychometric scale values can suggest more suitable scoring systems for data analysis – that is, a set of numeric scores that gives a better approximation to the perception of verbal labels. As our literature survey shows, custom scale values are very rarely used in current research, and only few methodological studies acknowledge this possibility (Labovitz, 1967, p. 155; Worcester & Burns, 1975, p. 191; see also Tukey, 1961, p. 246).

The paper starts out with a review of work published in a broad range of linguistic journals to examine the use of ordinal response scales in language research. The focus is on their structure (number of scale points and how they are verbalized), the underlying dimension that makes the categories ordinal (e.g., intensity, frequency), and how the data are analyzed statistically. Our survey points to three commonly used dimensions: agreement, intensity, and frequency. Section 3 reviews psychometric work on these dimensions, to map the metric properties of related scale point labels. The utility of these insights for the construction and evaluation of response scales is exemplified in Section 4. Section 5 then discusses the use of psychometrically grounded scoring systems for an interval-scale analysis of ordinal data. This dive into the heated and enduring controversy over "appropriate" statistics for ordinal data will carefully consider competing positions, both from a measurement-theoretic and a statistical viewpoint. The data-analytic perspective that emerges from this exercise emphasizes that the question of appropriateness concerns the interpretation of results rather than the statistical operations by which they have been arrived – a viewpoint that has in fact been expressed by a number of prominent (applied) statisticians (e.g., Anderson, 1961, p. 315; Mosteller, 1958, p. 288; Tukey, 1961, p. 246; Vellemann & Wilkinson, 1993). Section 6 presents a case study, where insights into the perception of quantifiers (*no-one, few, some, many, most, everyone*) are used to analyze data from a survey on morpho-syntactic language variation. Section 7 then closes with a summary.

2. Ordered response scales in language research

Let us start by examining the use of ordinal scales in the linguistic research literature. We included into our survey all articles published between 2012 and 2022 in 17 linguistic journals (4,441 publications in total), which range broadly across subfields and methodologies. For an overview, please refer to Web appendix 1 (https://osf.io/qfmh8). The search terms "rating scale", "rating task", "judg(e)ment task", "ordinal", "Likert", and "semantic differential" were used to extract potentially relevant documents (n = 909). We then manually identified those articles that employed an ordinal response scale (n = 405), where informants indicate some kind of assessment by choosing from an ordered set of categories. If a study relied on different scale formats[1], each layout entered our survey, yielding a greater number of response scales (n = 473) than articles in our database. A

[1] Differences had to occur along (one of) the features of main interest in our survey: number of response categories, the incorporation of verbal labels, and the underlying dimension.

tabular record with our coded data and background information on scale and study features forms part of the TROLLing post associated with this article (Sönning, 2024a).

As for their role in the research design, the ordinal scales in our survey were predominantly used to measure outcome (dependent) variables (n = 276; 58%), but also correlational (n = 44; 9%) and predictor (independent) variables (n = 65; 14%). Other purposes included sample description (n = 44; 9%) and stimulus validation (n = 44; 9%).

Table 1 gives an overview of the structure of graded scales and cross-tabulates (i) number of response categories, ranging from 3 to 11; and (ii) the way in which (verbal) information is incorporated into the scale, e.g., whether only the endpoints are labeled or whether the scale is fully verbalized. The figures in boldface give the overall distribution of these attributes. We note that it is quite typical to have 5 (37%) or 7 (27%) response categories. Excluding 64 scales whose layout could not be ascertained, we observe that in most cases (49%) only the endpoints of the scale are labeled. Our focus in the present study is on fully labeled sequences, where descriptors are given for all categories. In our survey, 173 instruments (42%) have this format. In Table 1, this subset is highlighted in grey.[2]

Table 1. Structure of ordinal scales used in rating tasks.

	Number of response categories										
Labels	3	4	5	6	7	8	9	10	11	Total	
Endpoints only		9	61	19	74	1	24	7	4	**199**	**49%**
Each category	12	29	80	33	14	1	4			**173**	**42%**
Other†		1	12	5	10		3	3	3	**37**	**9%**
No information		1	22	5	28	1	3	3	1	(64)	
Total	12	40	175	62	126	3	34	13	8	**473**	
	3%	**8%**	**37%**	**13%**	**27%**	**1%**	**7%**	**3%**	**2%**		

Note. †This category includes numbers-only (n = 20, 5%), endpoints-plus-midpoint (n = 15, 4%), midpoint-only (n = 1) and stars (n = 1).

These fully labeled scales can be grouped according to the underlying dimension that effects the rank order. We follow Rohrmann (2007) and identify five major dimensions, which are set out and exemplified in Figure 1; the illustrative examples are taken from the studies in our survey.

- Intensity: Intensifying adverbs denote the degree to which a certain attribute is present (e.g., *slightly*/*quite* acceptable).
- Agreement: The widely familiar Likert-type response format[3] (e.g., *strongly*/*mainly*/*somewhat disagree*), which often involves elements of intensification but forms a separate dimension due to its widespread use and bipolar nature.
- Frequency: Expressions denote the rate at which something happens (e.g., *rarely*, *frequently*).

[2] As an aside, it may be noted that empirical evidence suggests that adding labels to all response options (rather than only to the endpoints) improves the reliability and validity of rating scales (see Krosnick & Fabrigar, 1997, pp. 149-152 for a review). With regard to the numeric-conversion approach to data analysis, on the other hand, an advantage of the endpoint-only format is that it makes the use of a linear scoring system (i.e., equal numeric distances between categories) more defensible.

[3] The term "Likert scale" is often used to broadly refer to an ordinal response format. In its original sense (Likert, 1932), however, it denotes an aggregated (or summated) scale based on multiple items (see, e.g., Harpe, 2015). Each of these items, in turn, is measured using an ordinal rating scale, where respondents indicate their level of (dis)agreement to a statement.

- Probability: Phrases reflect the likelihood of some event (e.g., *unlikely*, *probable*).
- Quality: A 'good'-'bad' continuum, which may also draw on intensifiers but typically relies on different adjectives (e.g., *poor*, *fine*), making it a dimension in its own right.

Figure 1. The five principal dimensions used in graded scales: Examples.[4]

Let us examine how often these dimensions appear in our survey. The distribution of the 173 fully verbalized response scales, which is given in Table 2, shows that the most frequently employed scheme are agreement scales, which account for a third of the cases. Intensity features in 18% of the scales, followed by frequency (9%), probability and quality (each at 3%). The category "None/other" (34%) includes sequences that are fully labeled but do not represent an underlying perception, attitude, or belief. Table 2 also reports on the number of categories used in fully labeled scales, which hover between 3 and 9, with 5 response options (47%) being the most frequent layout.

Table 2. Distribution of fully verbalized scales by dimension, number of response categories and analysis strategy.

			Number of response categories							Analysis strategy†			
Dimension	N	%	3	4	5	6	7	8	9	Num	Des	Ord	Non
Agreement	57	**33%**		7	25	17	8			52	5		
Intensity	32	**18%**	1	10	16	3	2			23	4		1
Frequency	16	**9%**	2	3	9	2				13	3		
Probability	6	**3%**		2	2	2				4	2		
Quality	6	**3%**	1		4	1				5		1	
None/other	56	**33%**	8	7	24	8	4	1	4	45	7	3	1
Total	173		12	29	80	33	14	1	4	142	21	4	2
%			**7%**	**16%**	**47%**	**19%**	**8%**	**1%**	**2%**	**84%**	**12%**	**2%**	**1%**

Note. †Abbreviations: Numeric conversion, Descriptive statistics, Ordinal regression, Non-parameteric procedures (see text for details).

[4] All images in this paper have been published under the Creative Commons Attribution 4.0 license (CC BY 4.0, http://creativecommons.org/licenses/by/4.0) in the accompanying OSF project (https://osf.io/8f9q4/). All figures were drawn using the R packages 'lattice' (Sarkar, 2008) and 'ggplot2' (Wickham, 2016).

Next, we consider how data from fully verbalized scales are analyzed statistically.[5] Relevant tallies appear in the right-most columns in Table 2:

- An analysis was coded as employing *numeric conversion* (Num) if it relies on a scoring system to convert the ordered categories into numbers and then uses averages or ordinary (mixed-effects) regression.
- The label *descriptive statistics* (Des) denotes analyses that exclusively resort to 'appropriate' descriptive statistics (i.e., measures other than the mean, such as category counts/percentages or medians).
- Studies in the class *ordinal regression* (Ord) employ a form of categorical regression respecting the order of response levels.
- The category *non-parametric* (Non) refers to procedures for ranked observations (e.g., Wilcoxon signed-rank test).

Clearly, the numeric-conversion approach is the most popular analysis strategy; in 84% of the cases, a scoring system is used to summarize or model the data. Except for a single study, all analyses used a linear scoring system: By assigning running integers to the ordered responses (e.g., from 1 to 5 for a five-point scale), distances between categories are assumed to be evenly spaced.

We may summarize our findings on the use of fully verbalized scales as follows: They are a frequently employed scheme (42% of the scales in our survey, compared to 49% for endpoint-only layouts; see Table 1), and they usually consist of 4 to 6 categories. About a third of these are agreement scales, followed by intensity scales (18%); the other dimensions (frequency, probability and quality) are less prevalent. In the vast majority of cases (84%), the responses are analyzed as though they had been collected on a continuous scale; almost categorically, studies rely on a linear scoring system with equally-spaced numeric steps along the continuum.

3. Psychometric scale values of verbal response labels

Our literature survey has shown that fully labeled ordinal response scales in linguistic studies chiefly rely on three dimensions: agreement, intensity, and frequency. This section collects experimental results on the perceived meaning of associated scale point labels.[6] A summary of findings for the two less commonly used classes (probability, quality) are relegated to Web appendix 2 (https://osf.io/v9xph). The data produced by our survey can be found in the accompanying TROLLing post (Sönning, 2024a), and R code for reproducing the figures are deposited on the OSF (https://osf.io/wdvx6).

Rohrmann (2007) investigated the perception of labels that are used to indicate levels of agreement. Subjects performed different tasks, and we will concentrate on the "category scaling" one, where expressions had to be placed on an equally-apportioned 11-point scale. This scale is mapped to the [0,10] interval. Responses were collected from 164 participants, and the results we report here are those for all contexts combined (annoyance by noise, job satisfaction, context-free). Figure 2 shows the results from Rohrmann (2007):

- The dots reflect the average rating across informants (also reported numerically at the left margin of the graph).

[5] In two studies, two analysis strategies were applied to the data from the same ordinal scale, leading to multiple entries in this part of the table. In 6 studies, no analysis strategy could be ascertained. The totals therefore add to 169 entries (173 + 2 − 6).
[6] For similar research on German and Chinese scale point labels, see, e.g., Rohrmann (1978) and Au et al. (2011).

- The error bars extend ±1 standard deviation around the mean and reflect the variability of responses.

The set of agreement indicators studied by Rohrmann (2007) offers fine-grained resolution across the bi-polar spectrum. Terms denoting the neutral scale midpoint are judged consistently (e.g., *half-half*, *undecided*), as are labels at the extremities of the scale (*fully*/*strongly (dis)agree*).

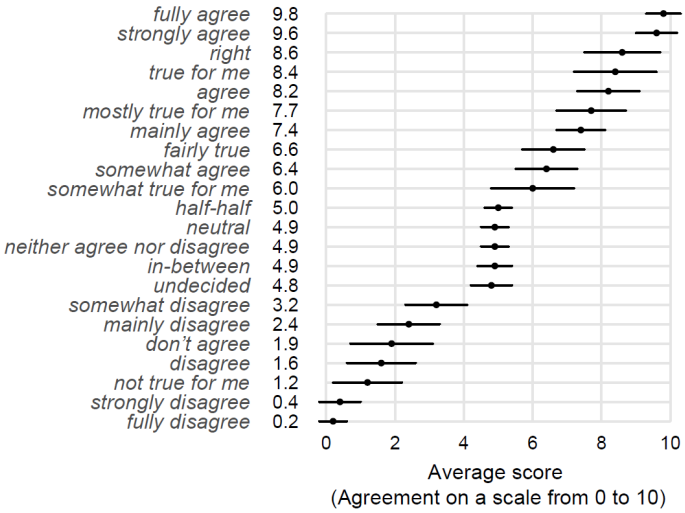


Figure 2. Scale values for the dimension agreement; data from Rohrmann (2007).

For intensity qualifiers, we summarize the findings of three studies that used similar methods to scale the meaning of intensifying adverbs (Krsacok, 2001; Matthews et al., 1978; Rohrmann, 2007). Informants were asked to locate each phrase on an 11-point scale, which we again map to the [0,10] interval. Figure 3 can be read as follows:

- The small grey dots indicate the ratings for four speaker groups; Krsacok (2001) studied two groups, male vs. female subjects.
- The black dots denote the average across these four groups, which is recorded at the left end of the graph

Again, the set of expressions yields fine increments across the scale. For intensifiers that were included in all three studies, the spread of the grey dots indicates the stability of perceptions – *mildly*, for instance, appears to be interpreted more consistently than *somewhat*.

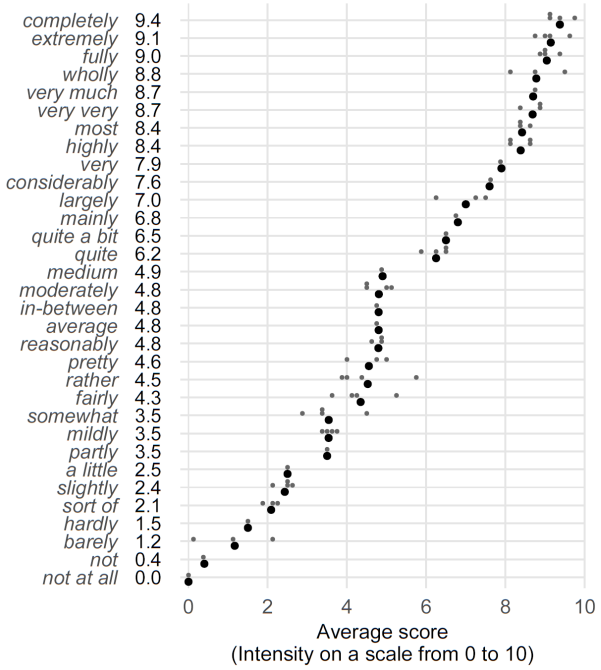


Figure 3. Scale values for the dimension intensity; data from Matthews et al. (1978), Krsacok (2001), and Rohrmann (2007).

Frequency expressions were studied by Mosteller and Youtz (1990). Using a mail questionnaire, they asked science writers to give the probability they would attach to 52 phrases. Ratings were obtained from around 230 individuals. Their findings are summarized in Figure 4:

- The filled black circles denote the average probability assigned to an expression by the 230 informants.
- The error bars extend from the lower to the upper quartile and therefore reflect variability across informants.
- The small grey dots show the distribution of estimates from 20 other studies, which are also summarized in Mosteller and Youtz (1990, p. 4), including the average probability observed in that study; thus, the single grey dot for *less often than not* shows that only Mosteller and Youtz (1990) studied this expression.
- The values reported at the left margin are averages across studies (weighted by the number of respondents).

We note that the frequency expressions in Figure 4 provide good coverage of the [0,100] interval. The consistency of interpretations varies, however. This can be seen from the error bars reflecting the spread of perceptions across individuals in Mosteller and Youtz (1990), and from the dispersion of averages across studies (small grey dots). For instance, the adverbs *sometimes* and *often*, which are popular anchors for graded scales, show considerable variation across studies, possibly indicating a lack of stability across contexts. Judging from the error bars, however, *often* seems to receive similar interpretations from individuals (compared to *sometimes* and nearby *usually*). Mosteller and Youtz (1990) offer a careful discussion of the observed variability across their 230 subjects and also asked respondents to indicate the range of percentages they consider acceptable for each expression. For *sometimes*, the lower and upper bounds were spread quite widely, on average, which points to a relatively fuzzy meaning of this frequency adverb.

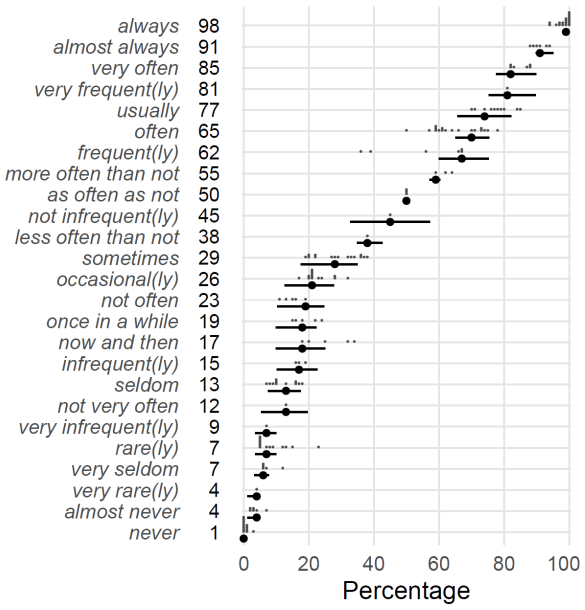


Figure 4. Scale values for frequency expressions; data from Mosteller and Youtz (1990).

Our review of psychometric work provides insight into the quantificational meaning of verbal scale point labels. We now look at how this information can be used in study design.

4. Use in study design

In this section, we illustrate how the experimental literature summarized above may aid in scale construction (Section 4.1) and demonstrate how psychometric findings may be used to evaluate the composition and statistical analysis of ordinal scales (Section 4.2).

4.1. Scale construction

When building ordinal response scales that are fully verbalized, it is generally preferable for labels to have meanings that divide up the continuum into approximately equal steps. If responses are expected to span the entire range of the scale, this will reduce measurement error (see Section 5.2.2). Further, the physical layout of a scale (i.e., the distance between tick boxes) typically suggests constant increments, and it has been observed that respondents' interpretations combine the semantics of the scale labels with their relative placement or position (Chase, 1969; Ironson & Smith, 1981; Klockars & Yamagishi, 1988; Schwarz et al., 1998).

We now use the scale values derived in Section 3 (overall averages reported at the left margin of Figures 2 and 3) to construct agreement and intensity scales that aim for roughly equal psychological intervals. Since previous research suggests that 5 to 7 response categories are optimal, we will concentrate on sequences of this length (for a review covering a number of criteria, see Krosnick & Fabrigar, 1997).

Figure 5a shows verbal anchors for 5-, 6- and 7-point agreement scales. These extend to the extremes of the dimension and are approximately equidistant. For intensity scales, Figure 5b gives three sets of labels. Here, scale design is complicated by the fact that attention must also be paid to collocation. Thus, the adverb *considerably* may not combine well with certain adjectives, in which case preference may have to be given to nearby alternatives (e.g., *very*).

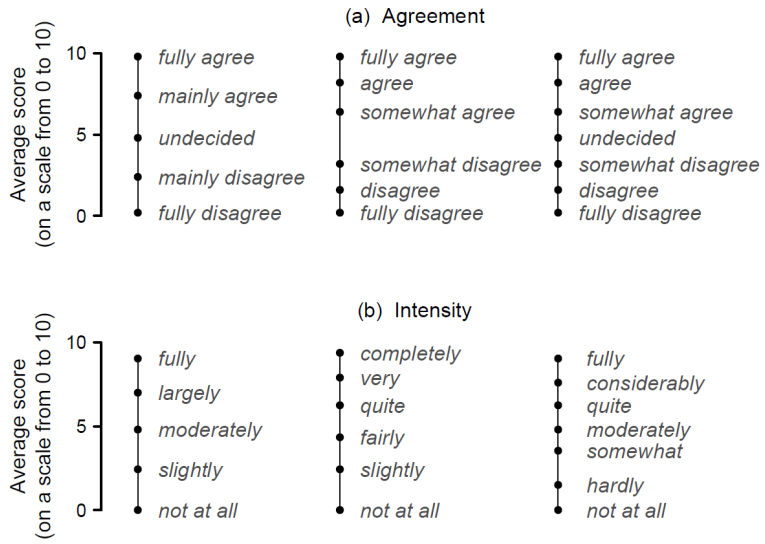


Figure 5. Scale construction: Approximately equal-interval scale labels for 5-, 6-, and 7-point (a) agreement and (b) intensity scales.

If responses are expected to cluster at one end of the scale, it may not be desirable to aim for equal increments. For instance, an acceptability study that deals with prescriptively ungrammatical constructions may choose verbal anchors that saturate the lower end of the acceptability spectrum, where most responses are expected to hover. Figure 6 illustrates two sets of intensifiers that enhance resolution at relatively high or low levels of intensity. Since it has been observed that respondents pay attention not only to the meaning of scale point labels but also their spatial arrangement, the physical layout of the scale may be adapted to (partly) preserve the unequal increments that are evident in Figure 6.

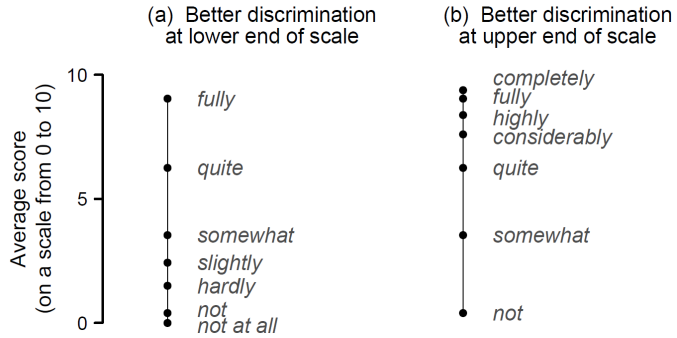


Figure 6. Scale construction: Saturation of the lower or higher end of the scale to enhance resolution in settings where ratings are expected (or known) to cluster near the extremes.

4.2. Scale evaluation

Psychological evidence may also be used to evaluate existing scales and their statistical treatment. To illustrate, let us consider three sets of frequency labels that we encountered in our survey. Figure 7 arranges these on the percentage scale. Set 1 increases in roughly equal steps and provides a good representation of the spectrum. In contrast, set 2 does not divide the continuum into proportionate

stretches – *sometimes*, the middle category, is closer in meaning to *seldom*. The equidistant numeric conversion used by the authors of that study therefore leads to distorted averages. Finally, set 3 saturates the lower end of the scale, where we find three tightly spaced anchors. The frequency expressions *occasionally* and *frequently*, on the other hand, are far apart – about half of the psychometric range covered by the instrument. The statistical analysis relied on a linear scoring system (numbers running from 1 to 6), which does not capture the perceived meaning of these expressions.

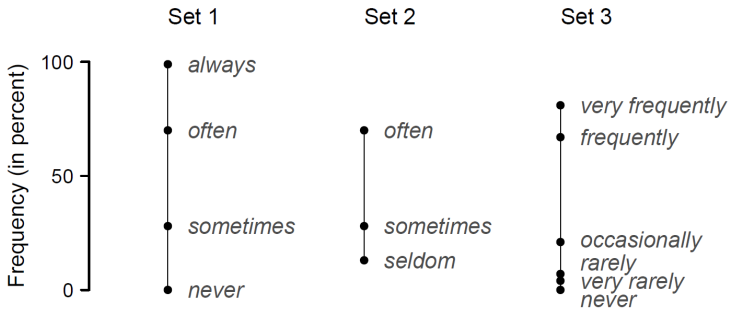


Figure 7. Scale evaluation: Saturation of the lower or higher end of the scale to enhance resolution in settings where responses are expected (or known) to cluster near the extremes.

This brings us to another way in which scaling information may inform empirical research: Instead of relying on a linear scoring system for an interval-scale analysis of ordinal variable (with equally-spaced integers, e.g., 0 to 6 for a 7-point scale), evidence on the quantificational meaning of scale point labels may be used to construct a more meaningful numeric translation. Despite their psychometric grounding, however, the use of custom scoring systems for the analysis of ordinal data must nevertheless proceed with caution. As we discuss in the next section, due consideration must be given to measurement-theoretic and statistical issues surrounding this approach to data analysis.

5. Use in data analysis

We now turn to a strategy on which opinions are divided: The use of numeric scores to analyze ordinal data. We start by recapitulating existing practices in language research (Section 5.1) and then look at the enduring controversy from a measurement-theoretic and a statistical perspective (Section 5.2). The aim is to make transparent deeper underlying issues, and to identify the implications they have for empirical work. In Section 5.3 we describe an informed data-analytic perspective that carefully negotiates the intersection between apparent benefits and inherent limitations.

5.1. Prevalence in linguistic research

Recall that our survey on the use of ordinal response scales in linguistic research revealed two striking facts: the pervasiveness of the numeric-conversion approach and the near-universal use of linear scoring systems to locate response categories on the number line.

To understand the majority practice by (language) scholars, let us consider the statistical options of an empirical researcher who has acquired ordinal data. We will compare alternative procedures with regard to two criteria: (i) informativity, by which we mean the nuance, or level of detail, they offer; and (ii) feasibility, which refers to the required know-how and software infrastructure. Figure 8 provides a schematic arrangement of techniques along these two dimensions. As a point of

reference, numeric-conversion strategies appear in grey: Arithmetic means are easy to compute and they allow for comparisons to be made at a fine level of detail. The same is true for ordinary linear (mixed-effects) regression, which may be categorized as a moderately sophisticated tool.

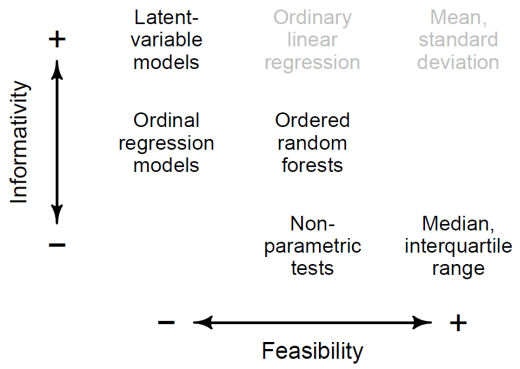


Figure 8. Procedures for summarizing the distribution of ordinal variables, arranged along two dimensions: informativity and feasibility.

Let us work our way from the bottom right (easy to implement but not very informative) to the top left (specialized procedures that bring into view fine-grained patterns).

- The median, though an "appropriate" measure of location for ordinal data, lacks nuance and fails to capture subtle differences between subgroups (see Krug & Sell, 2013, pp. 84-88).
- Non-parametric tests (e.g., the Wilcoxon signed rank test) require statistical software, offer limited flexibility and primarily yield inferential assessments; associated effect size measures lack interpretability (see Grissom & Kim, 2012, pp. 285-305).
- Random forests for ordinal outcomes call for specialized statistical software but require little (or no) user intervention. While they easily handle multifactorial data arrangements, methods of interpretation are currently limited to differences between predicted category probabilities (see Lechner & Okasa, 2022).
- Ordinal regression models (see Fullerton & Xu, 2017) are here classified as a relatively advanced technique, since they not only require statistical software but also rely on the user for model specification. Their default output, a regression table with thresholds and slopes on (say) the log odds scale, often proves difficult to interpret without appropriate post-processing steps.
- One type of ordinal regression, the cumulative-link model, has a mathematically equivalent latent-variable formulation (see Long, 1997, pp. 116-122; Agresti, 2010, pp. 53-55). This form of model yields fine-grained data summaries on a continuous scale; in terms of interpretability and informativity, it is therefore on a par with the numeric-conversion approach (see Sönning et al., 2024).

Figure 8 shows that the linguist who consciously decides against an interval-scale analysis will be forced to give up nuance, flexibility and/or ease of implementation, unless they resort to a procedure that is technically (much) more demanding and requires advanced statistical software. As this disincentivizes researchers from abandoning the numeric-conversion approach, this strategy will likely continue to permeate the empirical literature. This calls on methodologists to consider in more detail the arguments, viewpoints, and philosophies that have given rise to the continuing controversy surrounding the interval-scale analysis of ordinal data.

5.2. The controversy from three perspectives

To enable informed discussions and decisions about alternative approaches to analyzing rating scale data, it helps to look at underlying issues from a measurement-theoretic perspective (Section 5.2.1) and a statistical perspective (Section 5.2.2). This allows us to appreciate different positions and, more importantly, point out their implications and consequences for empirical work. These are summarized in what we will refer to as a data-analytic perspective (Section 5.2.3).

5.2.1. Representational vs. operational theories of measurement

To contextualize the debate on "appropriate" statistics for ordinal data, we must first recognize two different theories of measurement (see Hand, 1996; Knapp, 1990; Michell, 1986). For the empirical researcher, the key contrast between these frameworks is that they give different answers to the question of whether quantitative results reflect reality – i.e., whether patterns in the data can be interpreted at face value (see Michell (1986) for a lucid summary).

Stevens' (1946) taxonomy of scale types and "appropriate" statistics is firmly grounded in the representational theory of measurement. Briefly, representationalism is concerned with the mapping between numbers and real-world attributes, and with the question of which numbers preserve verifiable facts. When it comes to the choice of "appropriate" statistics, a key role is played by the notion of permissible transformations. These are changes to the numbers that are permissible in the sense that they still reflect ascertainable attributes of the objects of interest. For ordinal variables, the only confirmable information is their rank order. Any numeric representation that preserves the rank order of categories is therefore permissible. Categories can therefore be represented by any monotonically increasing set of scores, such as the integers 1-2-3-4 or 1-2-4-8. A statistical operation on these numbers (e.g., taking the arithmetic mean) is then considered "appropriate" if (and only if) the results it produces are stable across all permissible transformations.

Table 3 illustrates how a hypothetical group comparison based on arithmetic means depends on the scoring system: Using the set 1-2-3-4, group A has a higher mean; using the set 1-2-4-8, group B has a higher mean. The statistical conclusion based on the median, on the other hand, remains stable: Group A will always have a higher median. It is this invariance across permissible changes to the numerical representations that makes a statistic "appropriate". In exchange for this restriction, results can be interpreted as giving a one-to-one reflection of reality. They transcend the measurement instrument and allow the researcher to draw scale-free conclusions about the objects of study.

Table 3. Arithmetic mean vs. median: Dependence of directionality of group difference on the scoring system.

	Category frequencies (percentages)				Scoring system			
					1-2-3-4		1-2-4-8	
Group	I	II	III	IV	Mdn	Mean	Mdn	Mean
A	20%	20%	50%	10%	**3**	**2.5**	**4**	**3.4**
B	40%	20%	10%	30%	**2**	**2.3**	**2**	**3.6**

In contrast, operational measurement theory makes no reference to the real world. It defines the attribute of interest in terms of the (precisely specified) measuring procedure used, and thereby actively embraces the scale-dependence of data summaries. Since the number assigned to an observation emerges from a measurement operation, it is illogical to consider alternative numerical assignments. This renders the notion of permissible transformations irrelevant in this framework, which means that no scale-induced restrictions are imposed on the analysis. At the same time, however, this means that scale-free conclusions have no place in operationalism – results are not reflective of reality, which means that subject-matter interpretations are bounded by the measurement procedure. The empirical linguist adopting this philosophy must therefore exercise additional caution when drawing substantive conclusions from data.

Hand (1996) notes that these measurement theories map onto two different uses to which statistical models are often put (see Lehmann, 1990). Descriptive models, whose purpose is to summarize patterns in the data, are content with operationalism. Explanatory (or mechanistic/causal) models, which probe deeper into data-generating mechanisms and seek conclusions that go beyond the measurement scale, are aligned more closely with representationalism. The choice between measurement philosophies therefore also depends on the researcher's objectives.

5.2.2. Statistical reservations

We now turn to statistical limitations of the numeric-conversion approach and provide a brief summary of the main points (see Agresti, 2010, pp. 5–8, 137–140; Long, 1997, pp. 35–40, 116–119). These are numbered for cross-reference with Table 4, which provides an overview and summary. First, (1) the choice of scoring system may be ambiguous. To some extent, this point of criticism can be addressed by psychometric research, which may suggest sensible deviations from the near-universal use of linear scoring systems. Further, (2) response categories are usually consistent with a range of values on the underlying continuum. On the frequency dimension, for instance, a respondent's assessment may be "very often". If the response scale requires a choice between "often" and "always" (see Figure 7), the resulting score will be inflated or deflated by measurement error. A third (3) and perhaps minor point is that the researcher may obtain quantities of interest (predictions, estimates) that extend beyond the scale limits (e.g., a confidence interval stretching beyond "always").

The two final reservations, which arise from the boundedness of the scale, weigh more heavily since they systematically distort measures of location and spread. For one, (4) the variation of scores (as expressed, say, by the standard deviation) is downwardly biased near the endpoints of the scale. Subgroups whose responses gravitate towards the scale limits will therefore typically exhibit less variable ratings. For statistical inference, this may entail a violation of the homoscedasticity assumption (i.e., that the residual variation is approximately equal across conditions). Further, the assumption of normally distributed residuals, which underlies inferential procedures such as the *t*-test, ordinary regression, and ANOVA models, is untenable. This renders associated error probabilities (*p*-values and confidence intervals) dubious (see, e.g., Harwell & Gatti, 2001, pp. 111-112). Finally, (5) due to floor and ceiling effects, differences between arithmetic means will likewise be compressed near the bounds of any such scale (see Rohrer & Arslan (2021, pp. 5-6) for an illustration). In particular, this may affect the interpretation of interaction patterns, which may result from such local scale compressions (e.g., Loftus, 1978; Rohrer & Arslan, 2021).

Now that we have given due attention to measurement-theoretic and statistical issues, we are ready to formulate a data-analytic perspective for situations where psychometric research has shed light on the perceived distances between response categories.

5.2.3. A data-analytic perspective

Let us start by stating our position on the use of scoring systems for the analysis of ordinal data: If a researcher is able to rationalize and defend their choice of scores as yielding an approximate interval scale, and if they actively take into account the inherent limitations of the approach (i.e., the systematic distortedness and scale-dependence of their statistical conclusions), then they may choose to use numeric conversion as a pragmatic and informative analysis strategy. This view echoes the position of various prominent scholars (e.g., Anderson, 1961, p. 315; Mosteller, 1958, p. 288; Tukey, 1961, p. 246; Vellemann & Wilkinson, 1993) and Stevens (1946, p. 679) also gave a nod to this data-analytic perspective:

> "In the strictest propriety [...] means and standard deviations ought not to be used with these [i.e., ordinal] scales, for these statistics imply a knowledge of something more than the relative rank-order of data. On the other hand, for this 'illegal' statisticizing there can be invoked a kind of pragmatic sanction: In numerous instances it leads to fruitful results. While the outlawing of this procedure would probably serve no good purpose, it is proper to point out that means and standard deviations computed on an ordinal scale are in error to the extent that the successive intervals on the scale are unequal in size."[7]

Since a conscious awareness of the limitations we have summarized above is an essential component of this data-analytic perspective, let us recapitulate how different analysis strategies fare with regard to the statistical and measurement-theoretic issues outlined above. In Table 4, a filled circle indicates that a procedure fails to sidestep the drawback, hence: the fewer points, the better. We note that only "appropriate" statistics and ordered random forests derive scale-free conclusions from data. The results from numeric-conversion approaches, ordinal regression, and latent-variable models, on the other hand, are scale-dependent. As for the statistical reservations, ordinal regression and the latent-variable formulation of the cumulative-link model (along with ordered random forests and "appropriate" statistics) manage to sidestep all of the weaknesses listed above, which makes them attractive tools for ordinal data analysis. The simpler procedures that rely on scoring systems, on the other hand, suffer from these statistical drawbacks. Note that the use of psychologically grounded scoring systems addresses but one statistical criticism – the ambiguities involved in the choice of scores. Measurement error and boundary effects (i.e., non-constant variance and distortions due to compressions near the scale limits) continue to pose a threat.

Table 4. Statistical and measurement-theoretic drawbacks of different analysis procedures.

	Numeric conversion: Use of scoring system		Ordinal regression model†	Latent-variable model	Ordered random forest	"Appropriate" statistics
Limitations	Default	Custom				
Measurement-theoretic						
Scores do not reflect reality	●	●	●	●		
Statistical (see Section 5.2.2)						
(1) Choice of scores unclear	●					
(2) Measurement error	●	●				
(3) No hard scale bounds	●	●				
(4) Heteroscedasticity	●	●				
(5) Floor and ceiling effects	●	●				

Note. †This only includes ordered regression models other than the cumulative-link model, which is in fact mathematically equivalent to a latent-variable model.

[7] Here, Stevens reminds us that a linear scoring system may be a poor representation of the actual distances between categories, which is but one of the statistical concerns discussed above (i.e., point (1) in Table 4).

It follows that questions of (in)appropriateness should be directed at the subject-matter interpretations researchers advance on the basis of ordinal data (rather than the analysis strategy used). We now turn to a case study that relies on an interval-scale analysis of ordinal data. The purpose of this illustrative application is two-fold: First, we demonstrate that the choice between a default (linear) vs. a psychometrically grounded (custom) scoring system can affect the linguistic conclusions suggested by the data. Further, we carefully weigh statistical and measurement-theoretic limitations in light of the linguistic objectives underlying the study.

6. Illustrative application: Quantifier expressions in a survey on morphosyntactic variation

To illustrate how experimental research may inform the analysis and interpretation of rating scale data, we draw on the Bamberg Survey of Language Variation and Change (BSLVC; see Krug & Sell, 2013). We start by briefly sketching the research context (Section 6.1) and then review experimental work on the verbal labels used in the questionnaire (Section 6.2). Section 6.3 then analyzes the data using different scoring systems: an equidistant and a custom set of scores. The data are available from the TROLLing archive (Krug et al., 2024) and the R code for reproducing the analyses reported in this section can be found in the OSF (https://osf.io/5cfhe).

6.1. Background, research focus and questionnaire design

The BSLVC is a large-scale survey on the use of lexical and grammatical structures in different varieties of English. For details on the design and administration of the questionnaire, see Krug et al. (2024). We will turn to data from the grammar part, which asks respondents to indicate, on a 6-point scale, how prevalent a feature is in their home country or region; see Web appendix 3 for an illustration (https://osf.io/hzpqj). Specifically, participants give an estimate of how many speakers use the structure in question by choosing one of the following options: *no-one*, *few*, *some*, *many*, *most*, or *everyone*. These reported usage rates are elicited for two contexts, by asking informants to consider two settings: (i) an informal conversation among friends (informal speech); and (ii) an email to a former teacher (semi-formal writing). Accordingly, each sentence is presented in two modes (auditorily and in writing). For further considerations on the design of the BSLVC, see Krug and Sell (2013, pp. 79-84).

For our case study, we look at a single feature, the second person plural pronoun *yous(e)*, in two varieties: English and Scottish English. Data are available for 43 English informants (22 male, 21 female; mean age 21) and 61 Scottish informants (24 male, 37 female; mean age 21). Each participant provided two ratings (informal speech vs. semi-formal writing) for a single sentence (*Why don't youse come along to the restaurant?*). We are therefore looking at a mixed design, with one between-subjects factor (Variety: English vs. Scottish) and one within-subjects factor (Mode: speech vs. writing).

In general, it has been observed that *yous(e)* is more widespread in Scottish English (e.g., Smith 2012), and, since it classifies as non-standard, it is expected to be less prevalent in writing. We approach these data with the following questions: Is the usage rate of *yous(e)*, on average, higher (i) in Scottish English and (ii) in speech? For (i) and (ii), we expect affirmative answers from the data. Our third question, in contrast, is exploratory: (iii) Is there evidence for an interaction between Variety and Mode (i.e., does one variety show a detectably greater stylistic difference in usage rate?).

We start with a descriptive overview of the data. Figure 9 shows the distribution of the responses using a diverging bar chart (Heiberger & Robbins 2014); each stack of bars represents a condition

(Variety-Mode combination), and the bars are aligned at the scale midpoint, the boundary between *some* and *many*. Observed category percentages (i.e., the share of respondents who ticked a particular response option) are given to the right of the graph; for instance, 21% of the Scottish informants reported that "no-one" in their home country would use the sentence in an email to a former teacher (writing), compared to 64% of the English informants. We note that the prevalence judgments are, on average, higher for Scottish English, and for the spoken context. Whether the two varieties show similar stylistic clines is more difficult to see in the graph, and we address this question in Section 6.3 using a statistical model.

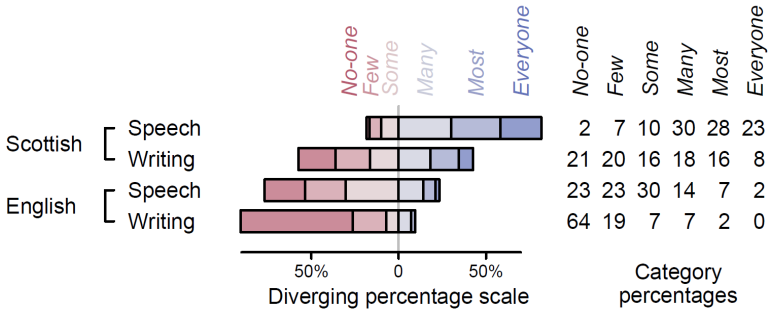
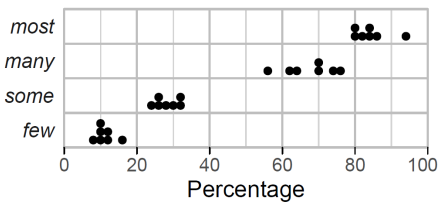


Figure 9. Distribution of ratings across responses categories, by variety (Scottish vs. English) and mode (speech vs. writing).

6.2. Quantifier perception: Survey of psychometric work

The continuum underlying this rating scale can be thought of as a relative frequency: *No-one* and *everybody* denote percentages of 0 and 100 and the perception of the intermediate quantifiers (*few*, *some*, *many*, *most*) can likewise be expressed as a proportion. Our literature search has returned three studies that use experimental techniques to probe how speakers interpret these quantifiers. A typical task asks subjects to consider a total of (say) 100 units – the set size – and to indicate how many units are, in their view, referenced by *some*. By averaging over multiple participants, the typical percentage denoted by a quantifier can then be estimated.[8] It is of interest to note that previous research has shown that the perceived proportional meaning of quantifiers varies systematically with set size: In smaller sets (e.g., 10 or 20 units), *few* and *some* are understood as referring to a greater proportional share (e.g., Newstead et al., 1987). Since the target set in our rating task is very large (i.e., a population of speakers), our summary of the literature will disregard set sizes smaller than 30.



[8] This procedure disregards the fact that the perception of quantifiers varies across individuals, and that the amount of between-speaker variation may vary among quantifiers. Nevertheless, these averages do provide a useful first approximation to the proportional meaning our informants are likely to attach to the expressions.

Figure 10. Survey on the perception of quantifier meanings (see Table 5 for details).

Proportional estimates reported in the literature for the quantifiers *few*, *some*, *many*, and *most* appear in Figure 10; information on the corresponding studies can be found in Table 5; for further details, please refer to Sönning (2024a). Estimates for the individual items show some variation across studies (and experiments within studies). At the bottom of Table 5, we summarize the percentages with a simple average across studies, and with a weighted average (printed in boldface), where figures are weighted in proportion to the number of subjects. Our scoring system will be based on these weighted means: *few* (11%), *some* (27%), *many* (67%), and *most* (83%). The two remaining scale labels, *no-one* and *everyone*, will be assigned values of 0% and 100%, respectively.

Table 5. Studies on the perception of quantifiers: Details.

Study	Subjects	Set size(s)	*few*	*some*	*many*	*most*
Borges & Sawyers 1974						
Experiment I	26	36, 48	11%	26%	57%	81%
Experiment II	30	36, 60, 84, 108	12%	32%	77%	94%
Newstead et al. 1987						
Experiment 1	234	108	11%	30%	70%	87%
Experiment 2	18	60, 108, 1000	16%	32%	74%	85%
Experiment 2 (additional data)	20	10000	9%	27%	70%	84%
van Tiel et al. 2021, Exp. 1a	165-489	432	10%	25%	62%	80%
Sönning 2024b	20	100	12%	28%	64%	83%
Mean						
Equally weighted			11%	28%	68%	85%
Weighted by sample size			**11**%	**27**%	**67**%	**83**%

The most notable difference to a default (linear) set of equidistant scores, then, is that the spacing between *some* and *many* is stretched, in accordance with experimental evidence on speakers' perceptions. Figure 11 illustrates how these custom scores are used to summarize and interpret the data; the illustrative ratings are those from the English informants for semi-formal writing. The bars denote the distribution of responses across the six categories; for instance, 28 informants reported that "no-one" in their home country would use this sentence in an email to a former teacher. If we replace each response with its custom (percentage) score, we can average over the 42 values (one participant failed to respond to this item in the written part of the questionnaire). This gives us a mean rating of 10.7%. We will interpret this average as an estimate of the usage rate of *yous(e)* in English semi-formal writing: According to the informants of our study, about 11% of the population in England would use *yous(e)* in semi-formal writing.

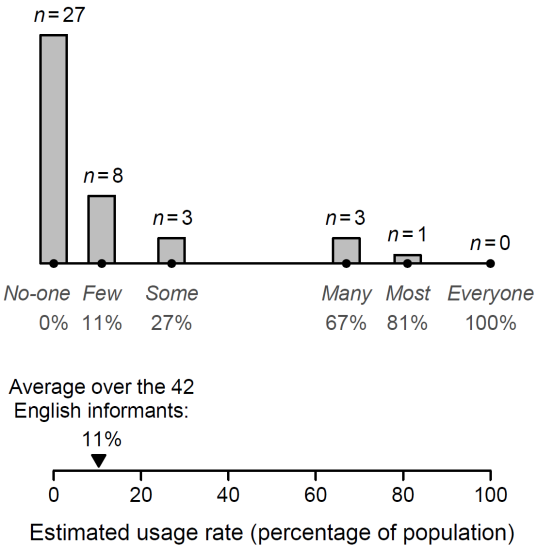


Figure 11. Illustration of the custom scoring system and its use for the analysis and interpretation of the data.

Similar to the dimensions frequency and probability, then, the psychometric scale values for the set of quantifiers provide a tangible frame of reference, as they are directly interpretable as relative frequencies (i.e., percentages). Accordingly, we will read averages on that scale as rough estimates of the overall prevalence of *yous(e)* in the varieties and contexts under investigation. As we will see further below, this will throw additional light on the scale-dependence of our statistical conclusions.

6.3. Numeric-conversion analysis

We will now analyze our data using two different scoring systems: a psychometrically grounded *custom scoring system*, and an equidistant *linear scoring system*. To make results more directly comparable, we will let the linear scoring system run from 0 to 100, in 20-point steps. It should be noted that the choice of scores (i.e., 0-20-40-60-80-100 rather than, say, integers running from 0 to 5) does not affect the insights emerging from the ensuing analyses. The two scoring systems are juxtaposed in Table 6.

Table 6. Scoring systems used in the analysis.

	Response category					
Scoring system	*No-one*	*Few*	*Some*	*Many*	*Most*	*Everyone*
Custom	0%	11%	27%	67%	81%	100%
Linear	0%	20%	40%	60%	80%	100%

The two factors (or independent variables) in our analysis, Variety (English vs. Scottish) and Mode (informal speech vs. semi-formal writing), are binary variables and crossed, making this as a 2x2 factorial design. Since we are dealing with a mixed design (with a between- and a within-subjects factor), we will analyze the data using a mixed-effects regression model that includes by-subject random intercepts.[9] The exploratory leg of our analysis probes into a potential interaction between

[9] The model was run using the lme4 package (Bates et al., 2015) in R (R Core Team, 2022), and specified as follows: lmer(rating ~ variety_c * mode_c + (1|subject)).

Variety and Mode – that is, whether the usage rate difference, or stylistic cline, between (informal) speech and (semi-formal) writing is comparable in the two varieties or not. A statistical interaction between the two factors would indicate that the stylistic clines reported by English and Scottish participants differ. We will adopt a widespread modeling strategy for this exploratory part of our analysis and rely on statistical criteria to decide whether the addition of an interaction term to our model finds support from the data.

Before we turn to regression tables and questions of model parsimony ("Occam's razor"), let us compare graphical model summaries (i.e., adjusted predictions) for the two scoring systems, based on a model that does include an interaction between Variety and Mode. These are shown in Figure 12. The estimates (adjusted predictions), which were obtained using the marginaleffects package (Arel-Bundock, 2023), are given numerically in Table 7, which also lists model-based comparisons (i.e., the reported stylistic cline for each variety and the difference between the English and Scottish clines). Before we consider these in more detail, note that the spacing of the quantifiers on the *y*-axes in Figure 12 differs: In panel (a), they are evenly spaced, and in panel (b) they are aligned with evidence on their interpretation, leading to a wider gap between *some* and *many*.

Figure 12 shows how the estimated usage rate varies by Variety and Mode: *Yous(e)* is used at a higher rate in speech and it is more prevalent in Scottish English. Panel (b) suggests that the stylistic cline is more pronounced in the Scottish subgroup: The difference between speech and writing is greater, which is reflected in the steeper slope – the lines fan out. For the linear scoring system, the estimated stylistic clines are more similar: 26.7 points for the Scottish informants and 20.1 points for the English informants. For the custom scoring system, the clines are wider apart: 29.9 vs. 17.3 points. Thus, panel (b) gives stronger indication of an interaction between Variety and Mode. Specifically, for the linear scoring system, the difference between the Scottish and English stylistic cline is estimated at 6.6 points vs. 12.6 points for the custom scoring system.

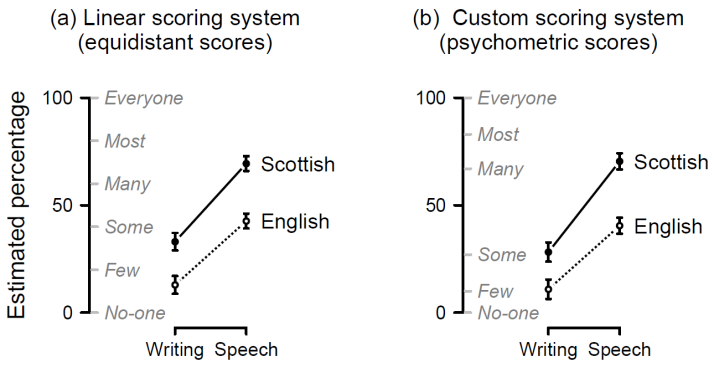


Figure 12. Model-based predictions for an analysis based on (a) default linear scoring system and a (b) psychometrically grounded scoring system; error bars denote standard errors (i.e., 68% confidence intervals).

Table 7. Model-based predictions and comparisons for the two scoring systems (standard errors in parentheses).

Analysis and variety	Speech	Writing	Stylistic cline (difference between speech and writing)	Difference between stylistic clines (difference between differences)
Linear scoring system				
English informants	33.0 (4.1)	12.9 (4.2)	20.1 (5.6)	6.6 (7.3)
Scottish informants	69.4 (3.5)	42.6 (3.4)	26.7 (4.7)	
Custom scoring system				
English informants	28.2 (4.5)	10.9 (4.5)	17.3 (6.1)	12.6 (8.0)
Scottish informants	70.4 (3.8)	40.5 (3.8)	29.9 (5.1)	

Let us now consider questions of model complexity and statistical inference. We have used the opportunity of a 2x2 factorial design to scale the regression terms in a way that enhances their interpretation (see Box et al., 2005, pp. 185-188).[10] The coefficient estimates for the two predictors, Variety and Mode, can be interpreted as average differences, which provide direct answers to questions (i) and (ii). This means that the coefficient for Variety gives the predicted difference between English and Scottish English, averaging over Mode (i.e., speech and writing). For the linear scoring system, this varietal difference, which was clearly notable in the diverging bar charts in Figure 9, is at 33.0 points (compared to 35.9 points for the custom scoring system). The difference between speech and writing (averaging over the two varieties), on the other hand, is at 23.4 points for model (a) and 23.6 points for model (b). Both scoring systems therefore yield differences in the expected direction. The standard errors denote the uncertainty surrounding these differences; values within two standard errors of an estimate mark an approximate 95% confidence interval. We observe that the average differences are estimated with sufficient precision, indicating statistically reliable patterns in the data.

Table 4. Regression coefficients and hypothesis test for an analysis† using a linear and a custom scoring system.

	Model (a) Linear scores		Model (b) Custom scores	
Coefficient	Estimate	(SE)	Estimate	(SE)
(Intercept)	39.5	(2.0)	37.5	(2.2)
Variety: Scottish – English	33.0	(4.0)	35.9	(4.4)
Mode: speech – writing (Δ)	23.4	(3.6)	23.6	(4.0)
Interaction: ΔScottish – ΔEnglish	6.6	(7.3)	12.6	(8.0)
Random intercept SD	8.2		8.9	
Residual SD	25.7		28.1	

Note. † Model specification: rating ~ variety_c * mode_c + (1|subject)

[10] To obtain the directly interpretable coefficients discussed shortly, all binary input variables are represented using contrast coding, with +0.5/−0.5 representing Scottish/English and speech/writing.

The coefficient for the interaction addresses question (iii): In model (a), the stylistic cline is 6.6 points steeper in Scottish English. This is the value we noted in the right-most column of Table 7 (the estimated difference between stylistic clines). Its standard error in model (a) indicates that, on inferential grounds, we may not be convinced that the interaction is necessary to capture the main patterns in the data. If we look at the interaction coefficient in model (b), however, we observe that it is appreciably greater, both in absolute terms (with 12.6 points about twice as large) and relative to its standard error.

This is also reflected in information criteria (IC), which assign greater predictive utility to the interaction term in model (b) vs. model (a).[11] This means that the statistical conclusions returned by our exploratory line of inquiry may very well depend on our choice of scoring system: Model (a) appears to favor the conclusion that the usage rate varies by Variety and Mode, and that the difference between speech and writing is roughly similar in the two populations. Model (b), on the other hand, is suggestive of a greater stylistic cline in Scottish English. Our analyses therefore reveal two things: First, when estimating average differences (questions (i) and (ii)), the choice of scoring system does not appear to matter much – both models observe similar differences between the varieties and usage contexts. However, if we go beyond overall trends and conduct subgroup analyses to examine interaction patterns (question (iii)), the choice of scoring system matters.

Our psychologically grounded choice of scores for the ordered categories therefore leaves us with a statistical interaction that we must now interpret in linguistic terms.

6.4. Scale-dependence of the interaction pattern

While our model offers some evidence for the presence of an interaction between Mode and Variety, the statistical reservations we discussed in Section 5.2.2 (see also Table 4) should make us cautious. After all, the difference in stylistic clines need not signal a "real" difference between the varieties – they could simply be due to floor effects that arise from the hard lower bound at zero. To get a second statistical opinion on the scale-dependence of this interaction, let us switch to an analysis strategy that removes boundary effects. We will redo the analysis using ordinal regression (a parallel cumulative model with a logit link) with the same fixed and random effects. This form of ordered regression has a (mathematically equivalent) latent-variable formulation (see Sönning et al., 2024), which facilitates interpretation and comparison with the patterns returned by our numeric-conversion analysis. Figure 13 shows that there is no statistical indication for an interaction on the latent-variable scale: The trend lines representing the stylistic gap between speech and writing are parallel. This suggests that the interaction is indeed scale-dependent.

[11] Keeping in mind that lower values are better, the addition of this term to the model should yield a lower IC score. Upon adding the interaction to the model, three information criteria (see Sonderegger, 2023, pp. 134-136, 279-280) show differential drops (Δ) in the two models. AIC: Δ = 4.7 in model (a) (1935.8 ⟶ 1931.1), Δ = 6.5 in model (b) (1973.4 ⟶ 1966.9); AICc: Δ = 4.5 in model (a) (1936.1 ⟶ 1931.6), Δ = 6.3 in model (b) (1973.7 ⟶ 1967.4); and BIC: Δ = 1.3 in model (a) (1952.4 ⟶ 1951.1), Δ = 3.2 in model (b) (1990.1 ⟶ 1986.9).

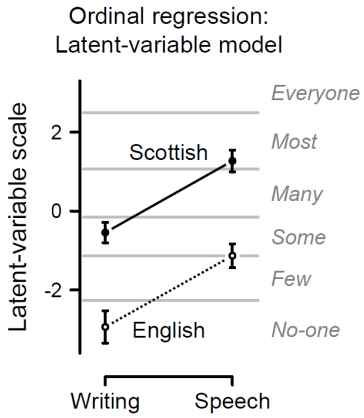


Figure 13. Model-based predictions for an ordered regression model on the latent-variable scale; error bars denote standard errors (i.e., 68% confidence intervals).

What implications does this have for the linguistic story we should be telling? It helps to call into mind the objectives of our study. The BSLVC is a large-scale questionnaire project that aims to profile and compare varieties of English with regard to a range of lexical and morpho-syntactic features. Based on informants' reports about the perceived usage rate of structures in their speech community, estimates are formed of their prevalence in varieties of English. By aiming for quantitative statements about populations of speakers, the goals of our analysis are descriptive rather than explanatory. This is because the primary purpose of the BSLVC is the documentation of cross-varietal and stylistic differences in the currency of non-standard and/or innovative features. Based on informants' reports, it provides estimates of how prevalent a particular feature is in a speech community (and stylistic context) – that is, how many speakers (expressed as a percentage) use it. No attempt is made (here) to explain or account for observed differences. Thus, we are not concerned with the question of *why* Scottish English may be showing a greater stylistic cline. From the viewpoint of measurement theory, then, we are free in our choice of scale, as we do not intend to attempt deeper causal interpretations.

After careful deliberation, we decide that the percentage scale (Figure 12) provides more informative answers to our research questions compared to a latent-variable scale (Figure 13): Percentages and percentage point differences are useful and interpretable indicators of (differences between) usage rates. This means that, given our applied purposes, the unbounded scale in Figure 13 offers no immediate interpretative advantage. Rather, the continuum on which we wish to describe and interpret our findings is naturally bounded: Floor and ceiling effects are real and we would be hesitant to remove them from the data summaries that form the basis of our linguistic interpretations. The statistical reservations that derive from boundary effects therefore lose some force, at least at the level of data description.

Nevertheless, several shortcomings of the ordinary mixed-effects regression model used above persist: (i) no allowance is made for measurement error due to the discreteness of the response categories; (ii) model-based quantities may very well extend beyond scale bounds; and (iii) statistical inferences may be unreliable due to violation of the normality and homoscedasticity assumption. For the data at hand, (ii) and (iii) could be partly addressed by modeling percentages on a non-linear scale such as log odds (using, say, fractional logit regression; Papke & Wooldridge, 1996). However, unless we make more fundamental changes to our analysis strategy (see Table 4), certain statistical reservations will continue to be a thorn in our side.

7. Summary and conclusion

The aim of the present paper was to illustrate how experimental findings may inform the construction of ordinal response scales and the arrangement of custom scoring systems. While the use of psychometric insights for scale design is uncontroversial and has been addressed in previous work, their usefulness for data analysis does not seem to enjoy widespread recognition. Seeing that the numeric-conversion approach to ordinal data is (and will likely remain) prevalent in language research, methodologists must embrace these realities and provide constructive advice that can cater for a vast majority of empirical researchers. A methodology grounded in actual research practice would arguably refrain from issuing overly firm recommendations on "appropriate" statistics and instead give balanced consideration to the tradeoffs involved in switching analysis strategies.

The fact that researchers almost exclusively rely on equal-spaced integers to analyze rating scale data is arguably unsatisfactory. To see moderate improvements in current practice, the use of custom scoring systems should be encouraged, perhaps drawing on the psychometric insights summarized above. More importantly, however, researchers must be made explicitly aware of the drawbacks and deceptions that may beset the interpretation of their results. Methodological advice should center on underlying measurement-theoretic and statistical arguments and clearly state their consequences for the analysis and interpretation of ordinal variables. The aim should be to foster an informed data-analytic discourse on this topic, and to delineate settings where researchers are particularly likely to be misled. A case in point is the study of interaction patterns, which may depend on the choice of scoring system and/or the analysis strategy used.

If such a grounded methodology were indeed to emerge, two directions of inquiry appear particularly fruitful. For one, the experimental evidence for the individual rating dimensions remains scarce. More psychometric work is needed on the measurement features of phrases, with a focus on the stability of interpretations across (populations of) subjects. By widening the empirical knowledge base, more informed recommendations for scale design and data analysis may be given. For instance, expressions that are open to a range of interpretations in the same population of speakers, or whose interpretation varies across populations, may be flagged as undesirable. Further, relatively unfamiliar expressions may also be unsuitable anchors on rating scales. Relevant work in this direction has been carried out by Mosteller and Youtz (1990) and Rohrmann (2007).

More systematic research is also needed on the scale-dependence and distortedness of descriptive and inferential results. This could involve sensitivity checks that look into the stability of conclusions across analysis strategies. A (crude) example of this line of inquiry is the comparison we drew with a latent-variable form of ordinal regression, a procedure that appeases statistical objections and thereby allows us to form some judgement on their effects on the scale-dependence of findings (for more thorough studies in this spirit, see Liddell & Kruschke, 2018; Rohrer & Arslan, 2021; Sönning et al., 2024). One of the key goals should be a set of heuristics that may be of help to practitioners who must decide on how much credence to lend to the output of a numeric-conversion analysis. This would shift the focus of discussion from the "appropriateness" of statistical procedures to where it is needed: on the linguistic interpretation of quantitative results.

Ultimately, this discourse must be cultivated in specialist fields and linguistically circumscribed areas of study. After all, the weight given to measurement-theoretic and statistical concerns will depend on the research context and the kinds of questions asked of the data. In experimental syntax, for instance, where interest centers on speakers' mental grammar, representationalist views on measurement cannot easily be dismissed. In applied research, on the other hand, the stated goals of a study may license a data-analytic mentality that takes advantage of custom scoring systems while giving careful attention to their limitations, to avoid drawing false conclusions from data.

Acknowledgements

I would like to thank Jan Vanhove and Manfred Krug as well as two anonymous reviewers for their constructive comments on an earlier version of this paper. Financial support from the German Research Foundation (DFG, grant number 548274092) is gratefully acknowledged.

References

Agresti, A. (2010). *Analysis of ordinal categorical data*. John Wiley & Sons.

Anderson, N. H. (1961). Scales and statistics: Parametric and nonparametric. *Psychological Bulletin, 58*(4), 305–316. https://doi.org/10.1037/h0042576

Arel-Bundock, V. (2023). *marginaleffects: Predictions, comparisons, slopes, marginal means, and hypothesis tests* (Version 0.13.0) [R package]. CRAN. https://CRAN.R-project.org/package=marginaleffects

Au, W., Rohrmann, B., Taylor, P., Ho, J. M., & Yeung, S. (2011). Developing equivalent Chinese and English scale-point labels for rating scales used in survey research. *Asian Journal of Social Psychology, 14*(2), 91-111. https://doi.org/10.1111/j.1467-839X.2010.01333.x

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

Beckstead, J. W. (2014). On measurements and their quality. Paper 4: Verbal anchors and the number of response options in rating scales. *International Journal of Nursing Studies, 51*, 807–814. https://doi.org/10.1016/j.ijnurstu.2013.10.002

Borges, M. A., & Sawyers, B. K. (1974). Common verbal quantifiers: Usage and interpretation. *Journal of Experimental Psychology, 102*(3), 335–338. https://doi.org/10.1037/h0035829

Box, G. E. P., Hunter, J. S., & Hunter, W. G. (2005). *Statistics for experimenters: Design, innovation, and discovery* (2nd ed.). John Wiley & Sons.

Chase, C. I. (1969). Often is where you find it. *American Psychologist, 24*, 1043. https://doi.org/10.1037/h0028186

Endresen, A., & Janda, L. A. (2016). 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.29915

Friedman, H. H., & Amoo, T. (1999). Rating the rating scales. *The Journal of Marketing Management, 9*(3), 114–123. https://doi.org/10.2139/ssrn.1652231

Fullerton, A. S., & Xu, J. (2017). *Ordered regression models: Parallel, partial, and non-parallel alternatives*. CRC Press.

Grissom, R. J., & Kim, J. J. (2012). *Effect sizes for research: Univariate and multivariate applications* (2nd ed.). Routledge. https://doi.org/10.4324/9780203802841

Hand, D. J. (1996). Statistics and the theory of measurement. *Journal of the Royal Statistical Society: Series A (Statistics in Society), 159*(3), 445–492. https://doi.org/10.2307/2983325

Harpe, S. E. (2015). How to analyze Likert and other rating scale data. *Currents in Pharmacy Teaching and Learning, 7*(6), 836–850. https://doi.org/10.1016/j.cptl.2015.08.001

Harwell, M. R., & Gatti, G. G. (2001). Rescaling ordinal data to interval data in educational research. *Review of Educational Research, 71*(1), 105–131. https://doi.org/10.3102/00346543071001105

Heiberger, R. M., & Robbins, N. B. (2014). Design of diverging stacked bar charts for Likert scales and other applications. *Journal of Statistical Software, 57*(5), 1–32. https://doi.org/10.18637/jss.v057.i05

Ironson, G. H., & Smith, P. C. (1981). Anchors away – the stability of meaning of anchors when their location is changed. *Personnel Psychology, 34*(2), 249–262. https://doi.org/10.1111/j.1744-6570.1981.tb00946.x

Jamieson, S. (2004). Likert scales: How to (ab)use them. *Medical Education, 38*(12), 1217–1218. https://doi.org/10.1111/j.1365-2929.2004.02012.x

Klockars, A. J., & Yamagishi, M. (1988). The influence of labels and positions in rating scales. *Journal of Educational Measurement, 25*(2), 85–96. https://doi.org/10.1111/j.1745-3984.1988.tb00296.x

Knapp, T. R. (1990). Treating ordinal scales as interval scales: An attempt to resolve the controversy. *Nursing Research, 39*(2), 121–123. https://doi.org/10.1097/00006199-199003000-00019

Krosnick, J. A., & Fabrigar, L. R. (1997). Designing rating scales for effective measurement in surveys. In L. Lyberg, P. Biemer, M. Collins, E. De Leeuw, C. Dippo, N. Schwarz, & D. Trewin (Eds.), *Survey measurement and process quality* (pp. 141–164). John Wiley & Sons.

Krsacok, S. J. (2001). Quantification of adverb intensifiers for use in ratings of acceptability, adequacy, and relative goodness [Doctoral dissertation, University of Dayton]. University of Dayton eCommons. https://ecommons.udayton.edu/graduate_theses/3652

Krug, M., & Sell, K. (2013). Designing and conducting interviews and questionnaires. In M. Krug & J. Schlüter (Eds.), *Research methods in language variation and change* (pp. 69–98). Cambridge University Press. https://doi.org/10.1017/CBO9781139419322.007

Krug, M., Schützler, O., Vetter, F., & Sönning, L. (2024). Background data for: The morpho-syntax of Scottish Standard English: Questionnaire-based insights. *DataverseNO*. https://doi.org/10.18710/B3NJBT

Labovitz, S. (1967). Some observations on measurement and statistics. *Social Forces, 46*(2), 151–160. https://doi.org/10.2307/2574436

Lechner, M., & Okasa, G. (2022). Random forest estimation of the ordered choice model. *arXiv*. https://doi.org/10.48550/arXiv.1907.02436

Lehmann, E. L. (1990). Model specification: The views of Fisher and Neyman, and later developments. *Statistical Science, 5*(2), 160–168. https://doi.org/10.1214/ss/1177012111

Likert, R. (1932). *A technique for the measurement of attitudes*. Columbia University Press.

Liddell, T. M., & Kruschke, J. K. (2018). Analyzing ordinal data with metric models: What could possibly go wrong? *Journal of Experimental Social Psychology, 79*, 328–348. https://doi.org/10.1016/j.jesp.2018.08.009

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

Long, J. S. (1997). *Regression models for categorical and limited dependent variables*. SAGE Publications. https://doi.org/10.4135/9781412984759

Matthews, J. J., Wright, C. E., Yudowitch, K. L., Geddie, J., & Palmer, R. L. (1978). The perceived favorableness of selected scale anchors and response alternatives. *U.S. Army Research Institute for the Behavioral and Social Sciences*.

Michell, J. (1986). Measurement scales and statistics: A clash of paradigms. *Psychological Bulletin, 100*(3), 398–407. https://doi.org/10.1037/0033-2909.100.3.398

Mosteller, F. (1958). The mystery of the missing corpus. *Psychometrika, 23*(4), 279–289. https://doi.org/10.1007/BF02288988

Mosteller, F., & Youtz, C. (1990). Quantifying probabilistic expressions. *Statistical Science, 5*(1), 2–34. https://doi.org/10.1214/ss/1177012258

Newstead, S. E., Pollard, P., & Riezebos, D. (1987). The effect of set size on the interpretation of quantifiers used in rating scales. *Applied Ergonomics, 18*(3), 178–182. https://doi.org/10.1016/0003-6870(87)90060-1

Norman, G. (2010). Likert scales, levels of measurement and the "laws" of statistics. *Advances in Health Sciences Education, 15*(5), 625–632. https://doi.org/10.1007/s10459-010-9222-y

Papke, L. E., & Wooldridge, J. M. (1996). Econometric methods for fractional response variables with an application to 401(K) plan participation rates. *Journal of Applied Econometrics, 11*(6), 619–632. https://doi.org/10.1002/(SICI)1099-1255(199611)11:6<619::AID-JAE418>3.0.CO;2-1

R Core Team. (2022). *R: A language and environment for statistical computing* (Version 4.2.0). R Foundation for Statistical Computing. https://www.R-project.org/

Rohrer, J. M., & Arslan, R. C. (2021). Precise answers to vague questions: Issues with interactions. *Advances in Methods and Practices in Psychological Science, 4*(2), 1–19. https://doi.org/10.1177/2515245920975120

Rohrmann, B. (1978). Empirische Studien zur Entwicklung von Antwortskalen für die sozialwissenschaftliche Forschung. *Zeitschrift für Sozialpsychologie, 9*, 222–245.

Rohrmann, B. (2007). Verbal qualifiers for rating scales: Sociolinguistic considerations and psychometric data. *University of Melbourne*. www.rohrmannresearch.net/pdfs/rohrmann-vqs-report.pdf

Sarkar, D. (2008). *Lattice: Multivariate data visualization with R*. Springer. https://doi.org/10.1007/978-0-387-75969-2

Schwarz, N., Grayson, C. E., & Knäuper, B. (1998). Formal features of rating scales and the interpretation of question meaning. *International Journal of Public Opinion Research, 10*(2), 177–183. https://doi.org/10.1093/ijpor/10.2.177

Smith, J. (2012). Scottish English and varieties of Scots. In B. Kortmann & K. Lunkenheimer (Eds.), *The Mouton world atlas of variation in English* (pp. 21–29). Walter de Gruyter.

Sonderegger, M. (2023). *Regression modeling for linguistic data*. MIT Press.

Sönning, L. (2024a). Background data for: Ordinal response scales: Psychometric grounding for design and analysis. *DataverseNO*. https://doi.org/10.18710/0VLSLW

Sönning, L. (2024b). Modeling the interpretation of quantifiers using beta regression. *Statistics for linguist(ic)s blog*. https://lsoenning.github.io/posts/2024-01-11_beta_regression_quantifiers/

Sönning, L., Krug, M., Vetter, F., Schmid, T., Leucht, A., & Messer, P. (2024). Latent-variable modelling of ordinal outcomes in language data analysis. *Journal of Quantitative Linguistics, 31*(2), 77-106. https://doi.org/10.1080/09296174.2024.1135129

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

Tukey, J. W. (1961). Data analysis and behavioral science or learning to bear the quantitative man's burden by shunning badmandments. In L. V. Jones (Ed.), *The collected works of John W. Tukey III* (pp. 187–389). Wadsworth.

van Tiel, B., Franke, M., & Sauerland, U. (2021). Probabilistic pragmatics explains gradience and focality in natural language quantification. *Proceedings of the National Academy of Sciences, 118*(9), e2005453118. https://doi.org/10.1073/pnas.2005453118

Velleman, P. F., & Wilkinson, L. (1993). Nominal, ordinal, interval, and ratio typologies are misleading. *The American Statistician, 47*(1), 65–72. https://doi.org/10.1080/00031305.1993.10475938

Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis* (2nd ed.). Springer. https://doi.org/10.1007/978-3-319-24277-4

Worcester, R. M., & Burns, T. R. (1975). A statistical examination of the relative precision of verbal scales. *Journal of the Market Research Society, 17*(3), 181–197.