Ordinal Data Analysis in R

Measuring Human Perceptions from Surveys

Matteo Ventura

2025-04-16

Description of the course

Surveys are key tools for measuring human perceptions, capturing latent traits through structured responses. Among the data they generate, ordinal and rating data are particularly important yet often less studied, requiring specialized statistical techniques. Ordinal data appears frequently in real-world applications, such as customer satisfaction surveys, psychological assessments, and medical research, making its correct analysis crucial for obtaining reliable insights. This short course provides instructor-led, hands-on training in the analysis of ordinal data. It begins with an overview of survey design and the validation of results, focusing on building effective surveys and ensuring the reliability of the data obtained. The course then covers the most commonly used statistical models for analyzing ordinal data, with an emphasis on discovering latent patterns and traits. Both theoretical foundations and practical applications will be explored, using real-world case studies from domains such as marketing, social sciences, tourism and culture.

A common approach to analyzing ordinal data is to treat it as numerical, but this can lead to a loss of statistical power. In this course, participants will learn how to apply specialized methods designed for ordinal data, allowing them to draw more effective and reliable conclusions.

Objectives of the course

By the end of the course, participants will have both theoretical knowledge and practical skills to analyze ordinal data in research and professional settings. Specifically, they will be able to:

- Understand what ordinal data is, how it differs from other types of data, and the challenges involved in its analysis
- Compute and interpret reliability and validity measures
- Fit proportional odds models in R and interpret the results
- Analyse rating data by applying CUB models

Introduction to Ordinal Data and Survey Design

The Role of Measurement in Science

Measurement is a fundamental activity in science, indeed we acquire knowledge about the world around us by observing it, and we usually quantify to give a sense to what we observe. Therefore, measurement is essential in a wide range of research contexts.

There exist several situations in which scientists come up with measurement problems, even though they are not interested primary in measurement. For instance:

1) A health psychologist needs a measurement scale which doesn't seem to exist. The study depends on a tool that can clearly distinguish between what individuals want to happen and what they expect to happen when visiting a physician. However, the review of previous research reveals that existing scales often blur this distinction, unintentionally mixing the two concepts. None of the available instruments capture the separation in the specific way her study requires. While the psychologist could create a few items that appear to address the difference between wants and expectations, she/he is concerned that these improvised questions may lack the reliability and validity necessary to serve as accurate measures.

2) An epidemiologist is conducting secondary analyses on data from a national health survey. They wish to investigate the link between perceived psychological stress and health status. Unfortunately, the survey did not include a validated stress measure. While it may be possible to construct one using existing items, a poorly constructed scale could lead to misleading conclusions.

3) A marketing team is struggling to design a campaign for a new line of high-end infant toys. Focus groups suggest that parents are heavily influenced by a toy's perceived educational value. The team hypothesizes that parents with strong educational and career aspirations for their children are more likely to be interested in the product. To test this idea across a broad, geographically diverse sample, the team needs a way to reliably measure parental aspirations. Something that additional focus groups can't easily provide.

Despite coming from different disciplines, these researchers share a common understanding: using arbitrary or poorly designed measurement tools increases the risk of collecting inaccurate data. As a result, developing their own carefully constructed measurement instruments appears to be the most reliable solution.

Historically, measurement problems were well-known in natural sciences such as physics and astronomy, even concerning figures like Isaac Newton. However, among social scientists, a debate arose regarding the measurability of psychological variables. While physical attributes like mass and length seem to possess an intrinsic mathematical structure similar to positive real numbers, the measurement of psychological variables was considered impossible by the Commission of the British Association for the Advancement of Science. The primary reason

was the difficulty in objectively ordering or summing sensory perceptions, as well illustrated by the question: how can one establish that a sensation of "a little warm" plus another similar sensation equals "twice as warm"?

Measurement classification

The americal psychologist Stevens (1946) disagreed with this perspective. He contended that the rigid requirement of "strict additivity," as seen in measurements of length or mass, was not essential for measuring sensations. He pointed out that individuals could make reasonably consistent ratio judgments regarding the loudness of sounds. For instance, they could determine if one sound was twice as loud as another.

Stevens further argued that this "ratio" characteristic enabled the data derived from such measurements to be mathematically analyzed. He is known for categorizing measurements into nominal, ordinal, interval, and ratio scales. In his view, judgments about sound "loudness" belonged to the ratio scale.

Despite the classification proposed by Stevens has been criticized by several authors and new classifications has been proposed, it is the most commonly accepted and used internationally.

Stevens identified four properties for describing the scales of measurement:

- **Identity**: each value has a unique meaning.
- **Magnitude**: the values of the variable have an ordered relationship to one another, so there is a specific order to the variables.
- **Equal intervals**: the data points along the scale are equally spaced, so the difference between data points one and two, is the same as data points three and four.
- **A minimum value of zero**: the scale has a true zero point.

As previously said, Stevens identified four scales of measurement, that is how variables are defined and categorised:

- **Nominal scale of measurement**: This scale has certain characteristics, but doesn't have any form of numerical meaning. The data can be placed into categories but can't be multiplied, divided, added or subtracted from one another. It's also not possible to measure the difference between data points. It defines only the identity property of data. Examples: Gender, Etnicity, Eye colour…
- **Ordinal scale of measurement**: It defines data that is placed in a specific order. While each value is ranked, there's no information that specifies what differentiates the categories from each other. These values can't be added to or subtracted from.
  Examples: satisfaction data points in a survey, where 'one = happy, two = neutral and three = unhappy.'

- **Interval scale of measurement**: The interval scale contains properties of nominal and ordered data, but the difference between data points can be quantified. This type of data shows both the order of the variables and the exact differences between the variables. They can be added to or subtracted from each other, but not multiplied or divided (For example, 40 degrees is not 20 degrees multiplied by two.).
  In this scale of measurement the zero is just a convention and not absolute, it is an existing value of the variable itself.
- **Ratio scale of measurement**: This scale include properties from all four scales of measurement. The data is nominal and defined by an identity, can be classified in order, contains intervals and can be broken down into exact value. Weight, height and distance are all examples of ratio variables. Data in the ratio scale can be added, subtracted, divided and multiplied. Ratio scales also differ from interval scales in that the scale has a 'true zero'. The number zero means that the data has no value point.
  An example of this is height or weight, as someone cannot be zero centimetres tall or weigh zero kilos.

Scales and Questionnaires development

Measurement plays a vital role across scientific disciplines, with each field creating specialized methods and tools tailored to its unique subjects of study. In the behavioral and social sciences, the area devoted to measurement is called psychometrics. This subfield concentrates on evaluating psychological and social constructs, which are most often assessed using questionnaires. Theaching how to build effective questionnaires would require a specific course, but this is out of the scope of this course. The following are some practical guidelines that researchers can use to develop measurement scales and questionnaires.

Step 1: Determine Clearly What It Is You Want to Measure

Researchers often discover their initial ideas about what they want to measure are vague, which can lead to costly changes later. Key questions include whether the scale should be theory-based or explore new directions, its level of specificity, and which aspects of the phenomenon to emphasize.

- **Define the theory**: Basing scale development on relevant substantive theories is essential for clearly defining the construct being measured, particularly when dealing with abstract or non-observable phenomena. A theoretical basis helps establish the construct's boundaries, reducing the risk of the scale extending into unrelated areas. In the absence of an existing theory, developers should create a conceptual framework of their own—beginning with a precise definition and linking the new construct to related, established ones.

- **Determine the level of specificity**: In psychometric scale development, it's important to consider how general or specific the measurement should be. This decision affects how well the scale works in predicting or relating to other variables. For example, if you're interested in general attitudes about personal control, a broad scale scale works well. But if you're studying beliefs about controlling a specific health issue, a focused scale is more appropriate.
- **Define which aspects are enphasised**: Scale developers must clearly distinguish the target construct from related ones. Scales can be broad (e.g., general anxiety) or narrow (e.g., test anxiety). Including items outside the intended focus can lead to confusion or inaccurate measurement. For example, in health contexts, physical symptoms caused by an illness might be mistaken for psychological symptoms (like depression), leading to misleading results. Therefore, item selection should match the specific research purpose and avoid overlap with unrelated constructs.

Step 2: Generate an Item Pool

When developing a psychometric scale, items should be **carefully selected** or created to match the specific construct you aim to measure. That means you need a clear idea of what the scale is supposed to do, and every item on the scale should reflect that goal.

Imagine the construct (like anxiety, motivation, or trust) as something hidden or latent, which can't be observed directly. The items on your scale are the visible signs or behaviors that reflect this hidden thing. So, each item acts like a small "test" of how much of that construct a person has. If your items truly measure the construct, then someone with a high level of the trait should tend to score higher on all of them.

When constructing the item pool, it is important to consider the following aspects:

- **The latent construct** A good scale includes multiple items to improve reliability, but every single item must still be strongly connected to the latent construct. You should think broadly and creatively when writing items to make sure they cover all the different ways the construct can be expressed—but without straying into measuring something else.
  A construct is a single, unified idea (like "attitudes toward punishing drug abusers") that can be thought of as causing how someone responds to related items. A category, on the other hand, is just a grouping of different constructs (like "attitudes" in general, or "barriers to compliance").
  Just because several items relate to the same category doesn't mean they measure the same underlying construct. For instance, "Barriers to compliance" is a category that can include many distinct things (fear of symptoms, cost concerns, distance to treatment, etc.). Each of these could be a separate construct with its own latent variable, so a scale that mixes these up wouldn't truly be unidimensional (i.e., measuring just one thing).

- **Redundancy** is crucial for reliability: multiple items allow common content to summate while idiosyncrasies cancel out. However, avoid superficial redundancy (e.g., minor wording changes, identical grammatical structures) which can inflate reliability estimates. Useful redundancy involves expressing the same core idea differently. Overly specific or redundant items within a broader scale can create subclusters (e.g., multiple specific anxiety items in a general emotion scale), potentially undermining unidimensionality and biasing the scale. This is less of a problem if the items match the scale's intended specificity.
- **The number of items** Start with more items than planned for the final scale (e.g., 3-4 times as many) to allow for careful selection and ensure good internal consistency. An initial pool 50% larger might suffice if items are hard to generate or fewer are needed for reliability. If the pool is too large, eliminate items based on criteria like lack of clarity or relevance.
- **The wording** Including both positively worded items (indicating the presence of the construct) and negatively worded items (indicating its absence or low levels) is a common strategy to reduce acquiescence bias—the tendency of respondents to agree with statements regardless of their content. However, reversing the wording can sometimes confuse participants, particularly in general population or community samples, and this confusion may reduce the scale's reliability.

Step 3: Determine the Format for Measurement

Numerous formats for questions exist. The researcher should consider early on what the format will be. This step should occur simultaneously with the generation of items so that the two are compatible.

Number of Response Categories: More options can increase variability, which is desirable, especially with fewer items. Continuous formats (like a thermometer scale) allow many gradations. However, too many options may exceed respondents' ability to discriminate meaningfully, leading to "false precision" and potentially increasing error variance. Avoid ambiguous wording (e.g., "several," "few") and confusing spatial arrangements. Consider the practicality of scoring very fine-grained responses.

Odd vs. Even Categories: For bipolar scales (e.g., agree/disagree), an odd number allows a neutral midpoint ("neither agree nor disagree," "not sure"), while an even number forces a choice toward one end. The choice depends on whether allowing neutrality is desirable or should be avoided.

Specific Formats:

Likert Scale: Presents declarative statements with response options indicating degree of agreement (e.g., strongly disagree to strongly agree). Options should represent roughly equal intervals. Often uses 5, 6, or 7 points, potentially including a neutral midpoint (though midpoint

wording can be subtle). Useful for opinions, beliefs, attitudes. Statements should generally be moderately forceful, aiming to elicit responses near the center of the scale from the "average" respondent to maximize variance and discrimination.

Semantic Differential: Presents a stimulus (e.g., "automobile salesmen") followed by pairs of opposite adjectives (e.g., honest/dishonest) separated by several response lines/spaces. Respondents mark the space reflecting their evaluation. Adjectives can be bipolar (friendly/hostile) or unipolar (friendly/not friendly). Sets of related adjective pairs can form a scale tapping an underlying variable (e.g., honesty).

Visual Analog Scale (VAS): Presents a continuous line between two descriptors (e.g., "No pain" to "Worst pain imaginable"). Respondents mark a point on the line. Allows continuous scoring but interpretation can be subjective, and comparisons across individuals may be difficult. Highly sensitive, making it useful for detecting subtle changes within individuals over time (e.g., pre/post intervention). May reduce bias from recalling previous discrete responses. Often used as single items, limiting internal consistency checks; multi-item VAS scales are preferable.

Numerical Response Formats & Neural Processes: Research suggests linear arrays of numbers may align with fundamental neural processing of quantity, potentially giving formats like Likert scales special merit.

Binary Options: Offer two choices (e.g., agree/disagree, yes/no, check if applies). Simple for respondents but yields minimal variability per item, requiring more items for adequate scale variance. The ease of response may allow for more items to be administered.

Item Time Frames: Consider the temporal aspect. Some scales imply an enduring trait (e.g., locus of control), while others assess transient states (e.g., depression "in the past week") or have separate state/trait forms (e.g., anxiety). The choice should be active and theory-driven, matching the nature of the phenomenon and the scale's intended use.

Step 4: Have Initial Item Pool Reviewed by Experts

Purpose: Experts knowledgeable in the content area review the item pool to maximize content validity. Tasks for Experts: Rate item relevance to the construct definition, helping confirm the definition and item-construct correspondence. Evaluate item clarity and conciseness, suggesting improvements to reduce ambiguity. Identify potential gaps or overlooked aspects of the phenomenon. Caution: Experts may not understand scale construction principles (e.g., the need for redundancy). The scale developer makes the final decisions based on informed judgment.

Step 5: Consider Inclusion of Validation Items

Rationale: Including additional items during development can aid later validation efforts. Types of Items: Detecting Problems: Items to assess response biases like social desirability.

Items correlating highly with social desirability may need exclusion. Standard scales (e.g., Strahan & Gerbasi; MMPI bias scales) can be included. Construct Validity: Measures of theoretically related (or unrelated) constructs can be included to examine convergent and discriminant validity early on.

Step 6: Administer Items to a Development Sample

Sample Size: Needs to be large enough to minimize subject variance as a concern and ensure stable item covariation patterns. Nunnally suggests 300, but smaller samples are sometimes used, depending on the number of items/scales. Risks of small samples include unstable results (e.g., inflated alpha estimates) and poor representation of the target population. Sample Representativeness: The sample should resemble the intended population. Non-representativeness can be quantitative (different mean level or range of the attribute) or qualitative (different relationships among items/constructs). Quantitative differences may be less problematic for assessing internal consistency. Qualitative differences, where items have different meanings or underlying structures in the sample versus the population (e.g., due to language/cultural differences), are more serious and can undermine the development effort.

Step 7: Evaluate the Items

Goal: Identify the best-performing items from the pool to form the final scale, assessing their relationship with the latent variable's true score.

Key Qualities & Analyses: High Intercorrelations: Items should correlate highly with each other, indicating they share a common latent variable and have higher individual reliability. Inspect the correlation matrix.

Reverse Scoring: Address negatively correlated items, potentially by reverse scoring them if they reflect the opposite pole of the construct. Reverse scoring can be done during administration (potentially confusing), data coding (tedious/error-prone), or electronically (easiest) using formulas like NEW = (k + 1) - OLD. If reverse scoring doesn't resolve inconsistent correlations, the item likely doesn't belong.

Item-Scale Correlations: Each item should correlate substantially with the sum of the other items (corrected item-scale correlation is preferred over uncorrected to avoid inflation).

Item Variances: Items should have relatively high variance, indicating they discriminate between individuals. Very low variance suggests poor discrimination. Markedly different variances might signal inconsistent error or violation of model assumptions (like essential tau equivalence).

Item Means: Means should be close to the center of the possible score range. Extreme means often lead to low variance and poor correlations. Check means/variances after initial selection based on correlations.

Dimensionality: Use factor analysis to determine if the items form a single, unidimensional set, which is an assumption for coefficient alpha.

Reliability (Alpha): Calculate coefficient alpha (or alternatives like omega if assumptions aren't met) to assess internal consistency – the proportion of variance due to the true score. Can be computed using statistical software (SPSS RELIABILITY, SAS PROC CORR ALPHA) or by hand using variance-based formulas (preferred) or correlation-based Spearman-Brown. Alpha ranges from 0 to 1 (negative alpha indicates problems like negative inter-item correlations). Common (subjective) benchmarks for research scales: <.60 unacceptable, .60-.65 undesirable, .65-.70 minimal, .70-.80 respectable, .80-.90 very good; >.90 consider shortening. Aim higher during development as alpha might drop in new samples. Scales for individual assessment (clinical, diagnostic) require much higher reliability (e.g., mid-.90s). Single-item measures cannot use alpha; test-retest is an imperfect alternative. Omega is an option if assumptions for alpha are unmet.

Step 8: Optimize Scale Length

Trade-off: Shorter scales reduce respondent burden, while longer scales are generally more reliable. The goal is an optimal balance. Brevity is pointless if reliability is too low. Consider shortening only when reliability is high. Dropping Items: Removing weak items (low correlation with others) can increase alpha, especially in shorter scales where each item has more impact. If an item's correlation is only slightly below average, keeping it usually benefits alpha more than removing it hurts. Process: Identify items for potential removal based on their impact on alpha (using software output), low item-scale correlations, or low squared multiple correlations (communality). Alpha Precision: Alpha itself is an estimate; its stability (reliability) increases with the number of items. Longer scales yield more consistent alpha values across administrations. Build in a safety margin, as alpha may decrease in new samples. Split Samples: If the development sample is large enough, split it (e.g., in half or unevenly). Develop/optimize the scale on one subsample and cross-validate (check stability of alpha and other stats) on the second subsample, whose data did not influence item selection. This helps assess if initial results were inflated by chance, though subsamples are still more similar than entirely separate samples.