---

title: "Ordinal Data Analysis in R"

subtitle: "Measuring Human Perceptions from Surveys"

# description: "This short course provides hands-on training in the statistical models for ordinal data analysis, emphasizing the discovery of latent patterns and more reliable conclusions than treating it as numerical."

author: "Matteo Ventura"

date: "04/16/2025"

draft: false

listing: true

categories: [ordinal data, CUB, survey]

format:

html:

toc: true

toc-depth: 2

toc-location: left

number-sections: true

theme: spacelab

highlight-style: tango

df-print: paged

css: styles.css

callout-appearance: simple

grid:

margin-width: 300px

body-width: 1000px

code-fold: true # <-- AGGIUNGI QUESTO

code-tools: true # <-- OPZIONALE: aggiunge pulsanti per mostrare/nascondere tutto

pdf:

latex-distribution: miktex # o tinytex se lo stai usando

documentclass: scrartcl

engine: lualatex

number-sections: true

execute:

echo: true

---

<!-- <p> -->

<!-- <a href="materials/CUM7vsCUB.r" download style="margin-right: 15px;"> -->

<!-- <i class="bi bi-file-earmark-code-fill" style="font-size: 1.2em; vertical-align: middle;"></i> -->

<!-- <span style="margin-left: 5px; vertical-align: middle;">R Script</span> -->

<!-- </a> -->

<!-- <a href="materials/CV\_VENTURA\_ITA\_nodata.pdf" download> -->

<!-- <i class="bi bi-file-earmark-slides-fill" style="font-size: 1.2em; vertical-align: middle;"></i> -->

<!-- <span style="margin-left: 5px; vertical-align: middle;">Slides</span> -->

<!-- </a> -->

<!-- </p> -->

::: {.cell}

<style> p { text-align: justify; } </style>

:::

<h1>Description of the course</h1>

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.

<h1>Objectives of the course</h1>

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

::: {.content-visible when-format="html"}

::: column-margin

<p>

<a href="materials/Script\_plots\_examples.R" download style="margin-right: 5px;">

<i class="bi bi-file-earmark-code-fill" style="font-size: 0.9em; vertical-align: middle;"></i>

<span style="margin-left: 2px; vertical-align: middle;">R Script</span>

</a>

</p>

<p>

<a href="materials/CV\_VENTURA\_ITA\_nodata.pdf" download>

<i class="bi bi-file-earmark-slides-fill" style="font-size: 0.9em; vertical-align: middle;"></i>

<span style="margin-left: 2px; vertical-align: middle;">Slides</span>

</a>

</p>

:::

:::

## 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 {.unnumbered .toc-ignore}

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.

### Determine Clearly What You Want to Measure {.unnumbered .toc-ignore}

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.

<!-- When building a scale, it's important to make sure it measures only the construct of interest—not other, similar ones. For instance, if you're interested in measuring test anxiety, including items about social anxiety would muddy the results. Similarly, if a depression scale includes physical symptoms like fatigue, it might confuse illness-related issues (like arthritis-related tiredness) with depression. This can distort results, especially in clinical or medical research. So, researchers need to be deliberate about item selection, ensuring the scale reflects exactly what they want to measure—and nothing else. -->

<!-- ::: {.callout-tip} -->

<!-- If a depression scale includes physical symptoms like fatigue, it might confuse illness-related issues (like arthritis-related tiredness) with depression. This can distort results, especially in clinical or medical research. -->

<!-- ::: -->

### Generate an Item Pool {.unnumbered .toc-ignore}

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.

<!-- Explanation: -->

::: {.callout-caution}

Reversing the wording of items (also known as reversed polarity) can confuse respondents, especially if the items are complex or abstract, ot if the respondents have lower reading comprehension or aren’t used to taking surveys.

This confusion can lead to inconsistent or inaccurate responses, which lowers the reliability of the scale (i.e., how consistently it measures the construct).

:::

<!-- When people respond to surveys, some tend to just agree with statements out of habit—this is called acquiescence bias. To prevent this, researchers often mix in negatively worded items (e.g., “I rarely feel confident” alongside “I feel confident”) to check whether the respondent is paying attention and to balance out response patterns. -->

<!-- However, this technique can backfire. Reversing the wording of items (also known as reversed polarity) can confuse respondents, especially if: -->

<!-- The items are complex or abstract. -->

<!-- The respondents have lower reading comprehension or aren’t used to taking surveys (as is sometimes the case in community or general population samples). -->

<!-- This confusion can lead to inconsistent or inaccurate responses, which lowers the reliability of the scale (i.e., how consistently it measures the construct). So, while including both types of items has its benefits, it should be done carefully, with attention to clarity and the characteristics of the target population. -->

<!-- Conclusion (Item Pool): The item pool should be large, relevant, and contain useful redundancy. Items should be grammatically sound, avoid ambiguity, and not force respondents into "package deals". -->

### Determine the Format for Measurement {.unnumbered .toc-ignore}

Defining the measurement format is a critical step in designing data collection instruments like questionnaires and scales. This decision, ideally made concurrently with item generation, impacts data quality, variability, instrument sensitivity, and ultimately, research conclusions.

<!-- Thurstone Scaling: Aims to create items that respond ("resonate") to specific levels of an attribute, like tuning forks. Judges typically sort items into piles representing equal intervals of the construct. Respondents' agreement patterns theoretically pinpoint their level of the attribute. Elegant but difficult to implement effectively. -->

<!-- Guttman Scaling: Items tap progressively higher levels of an attribute, so endorsing one implies endorsing all lower-level items. The score is the highest item endorsed. Works well for objective, hierarchical data (e.g., smoking frequency) but less so for subjective constructs where the order might vary across individuals. -->

<!-- Scales With Equally Weighted Items: Fit well with models where items are roughly equivalent indicators of a common phenomenon. Allows flexibility in response formats. -->

Most scale items consist of two parts: a stem and a series of response options. A kew aspect of the scale items is the number of response options. A desiderable quality of a measurement scale is variability.

One way to increase opportunities for variability is to include lots of scale items. Another way is to provide numerous respose options within each item, especially with fewer items.

In this view, continuous formats (e.g., thermometer scales) offer many gradations, and so increase the opportunities for variability. However, too many options can exceed respondents' ability to meaningfully discriminate, leading to "false precision" and increased error variance. Researchers must balance the need for variability with respondents' cognitive limitations.

Another issue the investigator has to concern with, is whether the number of options should be even or odd.

This choice depends on the type of question. the type of response option, and the objectives of the investigator.

An odd number of categories usually allows to express neutrality, while an even number of categories forces a choice from the respondent. The choice depends on whether allowing neutrality is desirable or should be avoided.

There exist several ways to present items that are commonly used:

- \*\*Likert scales\*\*: are widely used psychometric tools designed to measure attitudes, opinions, and perceptions by assessing the degree of agreement or disagreement with a statement. These scales typically present a statement (called a Likert item) followed by an ordered series of response options, generally consisting of five or seven points. However, scales with four, nine, or ten points can also be employed.

Response anchors are the labels that define each point on the scale (for example, "Strongly disagree," "Disagree," "Neutral," "Agree," "Strongly agree"). Scales with an odd number of points often include a neutral midpoint, while scales with an even number of points force the respondent to express a direction (agreement or disagreement).

Likert scales are extensively applied in surveys to assess employee engagement, customer satisfaction, product feedback, and clinical evaluations.

<!-- Key Insight: The structure of a Likert scale, including the number of points and the wording of the anchors, significantly influences the data collected. Researchers must carefully consider these factors to ensure the scale accurately measures the desired construct. -->

::: {.callout-warning}

Although Likert scale data is often numerically coded to facilitate analysis, it's essential to remember their ordinal nature and approach the calculation of means with caution.

:::

<!-- Median and mode represent more robust measures of central tendency for this type of data. -->

\* - \*\*Semantic Differential scales\*\*: are assessment tools used to measure attitudes and opinions toward an object, person, event, or idea through pairs of bipolar adjectives. Developed by psychologist Charles E. Osgood, these scales present a concept followed by several rows of opposite adjective pairs placed at the extremes of a continuum, typically with five to seven intermediate points.

Respondents evaluate the concept on each adjectival scale by selecting the point that best represents their attitude. Examples of bipolar adjective pairs include "Good - Bad," "Happy - Sad," "Strong - Weak," and "Pleasant - Unpleasant." These scales are commonly used in market research, branding, and customer satisfaction assessments to understand perceptions and associations.

Semantic differential scales explore the connotative meaning of a concept, revealing the emotional and evaluative dimensions of attitudes, unlike Likert scales which primarily focus on the degree of agreement.

<!-- Data visualization for semantic differential scales can be accomplished through semantic differential graphs (often resembling line profiles connecting the average ratings for each adjective pair) or by treating each scale as an ordinal variable and using divergent stacked bar charts. -->

<!-- Visualization Considerations: When visualizing semantic differential scale data, it's crucial to clearly represent the bipolar nature of the scales, allowing for easy comparison of responses across different adjective pairs for the same concept or between different groups for the same adjective pair. -->

- \*\*Rankings\*\*: represent data where items are ordered according to a specific criterion or preference. Respondents arrange items in a sequence based on their preference, importance, or another defined attribute. Examples include ranking favorite movies, product features by importance, or job candidates.

Ranking data indicates relative order but not the magnitude of difference between positions. The difference between the first and second positions might be substantial, while the difference between lower positions might be negligible.

<!-- Key Insight: Analysis of ranking data focuses on the position of items within the ordered sequence. Visualizations should emphasize the distribution of positions for each item and facilitate comparisons regarding how frequently items appear in different positions. -->

- \*\*Visual Analog Scale\*\* (VAS): Presents a continuous line between two descriptors and the respondents mark a point on the line. Therefore, it is clear that this scale allows continuous scoring but it has to be noted that interpretation can be subjective, and comparisons across individuals may be difficult. An advantage of this type of scale is that it is shghly sensitive, so it useful for detecting subtle changes within individuals over time; moreover they may reduce reduce bias from recalling previous discrete responses.

<!-- - \*\*Numerical Response Formats\*\*: 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). This type of option is simple for respondents but yields to minimal variability per item, therefore more items are required for obtaining an adequate scale variance. However, 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. -->

### Experts' review {.unnumbered .toc-ignore}

Expert review plays a key role in strengthening content validity during scale development. By drawing on their subject-matter expertise, reviewers help ensure that the items meaningfully represent the construct.

Experts are typically asked to assess how well each item reflects the construct definition, providing feedback that can confirm or refine the conceptual framework. They also evaluate the clarity and precision of item wording, offering suggestions to reduce ambiguity. In addition, experts may highlight important aspects of the construct that have been overlooked.

However, it's important to note that content experts may not be familiar with psychometric principles. For instance, they might recommend eliminating seemingly redundant items, not realizing that some redundancy is intentional and necessary for reliability. While expert input is highly valuable, final decisions should rest with the scale developer, who must balance expert judgment with methodological rigor.

<!-- ### Step 5: Include Validation Items {.unnumbered .toc-ignore} -->

<!-- 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 {.unnumbered .toc-ignore} -->

<!-- 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 {.unnumbered .toc-ignore} -->

<!-- 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 {.unnumbered .toc-ignore} -->

<!-- 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. -->

### Subsequent Steps in Scales Development {.unnumbered .toc-ignore}

Following the initial design of the questionnaire, including the selection and construction of appropriate scales, the next crucial phase involves preparing for validation and data collection. This includes strategically incorporating additional items aimed at facilitating later validation efforts, such as those designed to detect response biases or to assess the questionnaire's construct validity by measuring theoretically related concepts.

Subsequently, the questionnaire is administered to a development sample. It's essential that this sample is sufficiently large and representative of the target population to ensure stable results and minimize concerns about subject variance.

Once the data is collected, a thorough evaluation of the individual items is undertaken. This involves examining their intercorrelations to ensure they are measuring a common underlying construct, addressing any negatively correlated items through techniques like reverse scoring, and assessing the correlation of each item with the overall scale. Furthermore, the variance and means of the items are analyzed to ensure they discriminate effectively among respondents. Factor analysis is employed to confirm the dimensionality of the scale, and reliability, often measured by Cronbach's alpha, is calculated to assess the internal consistency of the items.

Finally, the length of the scale is optimized. This involves balancing the need for brevity to reduce respondent burden with the desire for higher reliability, which is generally associated with longer scales. Weak items that negatively impact reliability are considered for removal, and techniques like splitting the development sample for cross-validation can be used to ensure the stability of the optimized scale in new samples.

## Principles for Visualizing Ordinal Data

The most important principle in visualizing ordinal data is to always represent ordinal categories in their natural, ordered sequence in any visual representation. In bar charts, bars should be arranged along the axis based on the logical order of the ordinal scale (e.g., from "Low" to "High"). For stacked and divergent bar charts, the segments representing ordinal categories should also follow this intrinsic order within each bar.

```{r warning=FALSE}

library(ggplot2)

# Create sample data

satisfaction <- data.frame(

level = factor(c("Very Dissatisfied", "Dissatisfied", "Neutral", "Satisfied", "Very Satisfied"),

levels = c("Very Dissatisfied", "Dissatisfied", "Neutral", "Satisfied", "Very Satisfied")),

count = c(15, 23, 30, 45, 27)

)

# Create bar chart with ordered categories

ggplot(satisfaction, aes(x = level, y = count)) +

geom\_bar(stat = "identity", fill = "steelblue") +

theme\_minimal() +

labs(title = "Customer Satisfaction Levels",

x = "Satisfaction Level",

y = "Number of Responses") +

theme(axis.text.x = element\_text(angle = 45, hjust = 1))

```

The choice of chart should align with the research question and the specific aspect of ordinal data being investigated. Not all chart types are equally effective for representing ordered categorical data.

### Bar Charts {.unnumbered .toc-ignore}

Represent each ordinal category with a bar, whose height or length corresponds to the frequency or count of that category. Fundamentally, the bars must be arranged in the logical order of the ordinal variable (e.g., from lowest to highest category). They can be vertical or horizontal; horizontal orientation is often preferred for readability of long category labels.

Bar charts provide a clear and easily understandable visualization of the distribution of a single ordinal variable, highlighting the frequency of each ordered category.

```{r warning=FALSE, message=FALSE}

library(ggplot2)

library(dplyr)

# Create sample Likert scale data for one survey question

likert\_data <- data.frame(

response\_category = factor(

c("Strongly Disagree", "Disagree",

"Neutral", "Agree", "Strongly Agree"),

levels = c("Strongly Disagree", "Disagree",

"Neutral", "Agree", "Strongly Agree")

),

frequency = c(15, 27, 43, 85, 30)

)

# Create horizontal bar chart with properly ordered categories

ggplot(likert\_data, aes(x = response\_category,

y = frequency,

fill = response\_category)) +

geom\_bar(stat = "identity") +

scale\_fill\_manual(values = c(

"Strongly Disagree" = "#d7191c",

"Disagree" = "#fdae61",

"Neutral" = "#ffffbf",

"Agree" = "#abd9e9",

"Strongly Agree" = "#2c7bb6"

)) +

coord\_flip() + # Horizontal orientation for better label readability

theme\_minimal() +

labs(

title = "Responses to: 'The new software

interface is intuitive to use'",

subtitle = "Distribution of 200 employee responses",

x = "",

y = "Number of Responses"

) +

theme(

legend.position = "none", # Remove legend as colors are self-explanatory

axis.text.y = element\_text(size = 12),

plot.title = element\_text(face = "bold"),

panel.grid.major.y = element\_blank() # Remove horizontal grid lines

)

```

### Stacked Bar Charts {.unnumbered .toc-ignore}

Show multiple ordinal categories within a single bar, with each segment representing a different category stacked on top of another. They are useful for comparing the distribution of ordinal data across different groups or conditions. They can be displayed as counts or as percentages (where each bar totals 100%).

Stacked bar charts allow comparison of both total amounts within each group and the proportion of each ordinal category within those groups, providing insights into how distributions differ between categories.

```{r warning=FALSE}

library(ggplot2)

library(dplyr)

library(tidyr)

# Create sample semantic differential scale data

# This represents evaluations of three different smartphones on five dimensions

semantic\_data <- data.frame(

product = rep(c("Smartphone A", "Smartphone B", "Smartphone C"), each = 5),

dimension = rep(c("Ineffective - Effective",

"Complicated - Simple",

"Unreliable - Reliable",

"Outdated - Innovative",

"Unattractive - Attractive"), 3),

rating\_1 = c(5, 8, 3, 10, 7, # Smartphone A

2, 5, 2, 3, 4, # Smartphone B

8, 6, 4, 5, 3), # Smartphone C

rating\_2 = c(10, 12, 15, 8, 13, # Smartphone A

8, 10, 7, 9, 11, # Smartphone B

12, 8, 9, 10, 7), # Smartphone C

rating\_3 = c(25, 20, 22, 18, 20, # Smartphone A

15, 20, 18, 13, 15, # Smartphone B

20, 22, 18, 15, 20), # Smartphone C

rating\_4 = c(35, 30, 32, 40, 35, # Smartphone A

45, 35, 40, 38, 35, # Smartphone B

30, 35, 38, 40, 35), # Smartphone C

rating\_5 = c(25, 30, 28, 24, 25, # Smartphone A

30, 30, 33, 37, 35, # Smartphone B

30, 29, 31, 30, 35) # Smartphone C

)

# Reshape data for ggplot

semantic\_long <- semantic\_data %>%

pivot\_longer(cols = starts\_with("rating\_"),

names\_to = "rating\_level",

values\_to = "count") %>%

mutate(

rating\_number = as.numeric(substr(rating\_level, 8, 8)),

rating\_label = factor(

case\_when(

rating\_number == 1 ~ "1 (Negative)",

rating\_number == 2 ~ "2",

rating\_number == 3 ~ "3 (Neutral)",

rating\_number == 4 ~ "4",

rating\_number == 5 ~ "5 (Positive)"

),

levels = c("1 (Negative)", "2", "3 (Neutral)", "4", "5 (Positive)")

)

)

# Calculate percentages for each product-dimension combination

semantic\_pct <- semantic\_long %>%

group\_by(product, dimension) %>%

mutate(

percentage = count / sum(count) \* 100,

total = sum(count)

) %>%

ungroup()

# Create stacked bar chart

ggplot(semantic\_pct, aes(x = dimension, y = percentage, fill = rating\_label)) +

geom\_bar(stat = "identity") +

facet\_wrap(~ product) +

scale\_fill\_manual(values = c("1 (Negative)" = "#d7191c",

"2" = "#fdae61",

"3 (Neutral)" = "#ffffbf",

"4" = "#a6d96a",

"5 (Positive)" = "#1a9641")) +

coord\_flip() +

theme\_minimal() +

labs(title = "Smartphone Evaluations using Semantic Differential Scales",

subtitle = "Distribution of ratings across five dimensions",

x = "",

y = "Percentage of Responses",

fill = "Rating") +

theme(

legend.position = "bottom",

strip.background = element\_rect(fill = "lightblue", color = NA),

strip.text = element\_text(face = "bold")

)

```

This visualization effectively reveals patterns such as which smartphone is perceived as more innovative, which has the most consistent ratings across dimensions, and where the greatest differences between products exist. These insights would be difficult to discern from tables of raw data.

The stacked bar format is particularly effective for semantic differential scales because it shows the full distribution of responses, not just averages, allowing you to see whether opinions are polarized or consistent across respondents.

### Divergent Stacked Bar Charts {.unnumbered .toc-ignore}

Specifically designed to visualize ordinal data with a neutral central category or bipolar responses, such as Likert scales and semantic differentials. Segments representing responses on one side of the neutral point extend in one direction, while segments representing responses on the other side extend in the opposite direction from a central baseline. They effectively illustrate the balance between positive and negative responses and the distribution of opinions.

Divergent stacked bar charts are the recommended visualization for Likert-type scales as they clearly show the proportion of responses in each category and the overall tendency of agreement or disagreement.

```{r warning=FALSE}

library(ggplot2)

library(tidyr)

library(dplyr)

# Create sample Likert scale data

likert\_data <- data.frame(

question = c("The product is easy to use",

"The customer service was helpful",

"I would recommend this product to others"),

strongly\_disagree = c(5, 10, 8),

disagree = c(15, 18, 12),

neutral = c(20, 25, 15),

agree = c(40, 30, 35),

strongly\_agree = c(20, 17, 30)

)

# Reshape data for ggplot

likert\_long <- likert\_data %>%

pivot\_longer(cols = -question,

names\_to = "response",

values\_to = "count") %>%

mutate(response = factor(response,

levels = c("strongly\_disagree", "disagree", "neutral",

"agree", "strongly\_agree")),

response\_type = ifelse(response %in% c("strongly\_disagree", "disagree"),

"negative",

ifelse(response == "neutral", "neutral", "positive")),

# Negative values for disagreement

plot\_value = ifelse(response\_type == "negative", -count, count),

# For neutral, split half to each side (as per original code)

plot\_value = ifelse(response\_type == "neutral", count/2, plot\_value))

# Create divergent stacked bar chart

ggplot(likert\_long, aes(x = question, y = plot\_value, fill = response)) +

geom\_bar(stat = "identity", position = "stack") +

scale\_fill\_manual(values = c("strongly\_disagree" = "darkred",

"disagree" = "lightcoral",

"neutral" = "lightgrey",

"agree" = "lightblue",

"strongly\_agree" = "darkblue")) +

coord\_flip() +

theme\_minimal() +

labs(title = "Responses to Product Survey",

x = "",

y = "Count",

fill = "Response") +

theme(legend.position = "bottom")

```

### Other Possible Visualizations {.unnumbered .toc-ignore}

Depending on the specific analytical objective, these alternative visualizations can provide valuable perspectives on ordinal data, particularly when exploring relationships between variables or tracking changes in rankings.

\*\*Mosaic plots\*\* show the relationship between two or more categorical variables, including ordinal ones, using tiled rectangles whose area is proportional to the frequency of each combination of categories.

```{r warning=FALSE}

# install.packages("ggmosaic")

library(ggplot2)

library(ggmosaic)

library(dplyr)

# Create sample data for education level (ordinal)

#and job satisfaction (ordinal)

set.seed(123)

n <- 500

education\_levels <- c("High School", "Associate's", "Bachelor's",

"Master's", "Doctorate")

satisfaction\_levels <- c("Very Dissatisfied", "Dissatisfied",

"Neutral", "Satisfied", "Very Satisfied")

# Create sample data with a pattern

#(higher education tends to correlate with higher satisfaction)

mosaic\_data <- data.frame(

education = factor(sample(education\_levels, n, replace = TRUE,

prob = c(0.3, 0.25, 0.25, 0.15, 0.05)),

levels = education\_levels),

satisfaction = factor(NA, levels = satisfaction\_levels)

)

# Generate satisfaction levels with

#some correlation to education

for (i in 1:n) {

# Higher education levels tend to have higher satisfaction probabilities

edu\_level <- which(education\_levels == mosaic\_data$education[i])

# Adjust probabilities based on education level

probs <- c(0.25, 0.25, 0.2, 0.2, 0.1) # Base probabilities

# Shift probabilities based on education level

shift <- (edu\_level - 3) \* 0.05 # Shift factor based on education

# Adjust probabilities (higher education gets

#more weight for higher satisfaction)

adjusted\_probs <- probs + c(-0.1, -0.05, 0, 0.05, 0.1) \* edu\_level

# Ensure probabilities are valid

adjusted\_probs <- pmax(adjusted\_probs, 0.01)

adjusted\_probs <- adjusted\_probs / sum(adjusted\_probs)

mosaic\_data$satisfaction[i] <- sample(satisfaction\_levels, 1, prob = adjusted\_probs)

}

# Create mosaic plot

ggplot(data = mosaic\_data) +

geom\_mosaic(aes(x = product(education),

fill = satisfaction)) +

scale\_fill\_brewer(palette = "RdYlGn", direction = 1) +

labs(title = "Relationship Between

Education Level and Job Satisfaction",

subtitle = "Mosaic plot showing the

distribution of satisfaction within

each education level",

x = "Education Level",

y = "Job Satisfaction",

fill = "Satisfaction Level") +

theme\_minimal() +

theme(axis.text.x = element\_text(angle = 45, hjust = 1),

legend.position = "bottom")

```

This mosaic plot visualizes the relationship between two ordinal variables: education level and job satisfaction. The width of each column represents the proportion of respondents with that education level in the overall sample. Within each education level column, the height of each colored section represents the proportion of respondents reporting that satisfaction level.

\*\*Line charts\*\* (bump charts) visualize the change in rank of different items over time or between categories, emphasizing movement in relative positions.

```{r warning=FALSE}

library(ggplot2)

library(tidyr)

library(dplyr)

# Create sample data for product rankings over time

rankings <- data.frame(

product = rep(c("Product A", "Product B",

"Product C", "Product D", "Product E"), 4),

quarter = rep(c("Q1 2024", "Q2 2024",

"Q3 2024", "Q4 2024"), each = 5),

rank = c(1, 2, 3, 4, 5, # Q1 rankings

1, 3, 2, 5, 4, # Q2 rankings

2, 1, 3, 5, 4, # Q3 rankings

3, 1, 2, 4, 5) # Q4 rankings

)

# Create bump chart

ggplot(rankings, aes(x = quarter, y = rank,

group = product, color = product)) +

geom\_line(size = 1.5) +

geom\_point(size = 4) +

scale\_y\_reverse(breaks = 1:5) + # Reverse Y-axis so

# rank 1 is at the top

theme\_minimal() +

labs(title = "Product Rankings by Quarter",

subtitle = "Showing changes in ranking

position over time",

x = "Quarter",

y = "Rank (Lower is Better)",

color = "Product") +

theme(legend.position = "bottom")

```