Machine Learning for Psychophysical Scaling with Ordinal Comparisons

Dissertation

der Mathematisch-Naturwissenschaftlichen Fakultät

der Eberhard Karls Universität Tübingen

zur Erlangung des Grades eines

Doktors der Naturwissenschaften

(Dr. rer. nat.)

vorgelegt von

David-Elias Künstle

aus Reutlingen

Tübingen

2024

Gedruckt mit Genehmigung der Mathematisch-Naturwissenschaftlichen
Fakultät der Eberhard Karls Universität Tübingen.

Tag der mündlichen Qualifikation:	01.10.2024
Dekan:	Prof. Dr. Thilo Stehle
1. Berichterstatter/-in:	Prof. Felix A. Wichmann, D.Phil.
2. Berichterstatter/-in:	Prof. Thomas S. A. Wallis, Ph.D.

Abstract

Objective measurement methods of subjective stimulus intensity have occupied scientists for over 100 years. The latest generation of these so-called psychophysical scaling methods combines experimental tasks in which subjects compare the similarity of stimuli with ordinal embedding algorithms developed in machine learning to obtain robust and multidimensional point representations of stimulus perception. However, even for experts, the correct application of ordinal embedding based scaling methods is technically and methodologically challenging.

In this dissertation, I describe a pipeline to make psychophysical scaling with ordinal comparisons more accessible using machine learning techniques. First, I introduce an open-source Python toolbox. This toolbox provides the most important algorithms and methods as user-friendly and efficient implementations, making ordinal embedding methods more accessible to psychophysicists. I then develop a procedure to simplify the essential choice of scale dimensionality based on statistical considerations and propose analysis methods to estimate the quality and variability of a scale to draw scientific conclusions. At last, I present a novel application of comparison-based scaling methods in a virtual reality experiment to measure distortions of varifocal glasses that lead to serious side effects such as dizziness.

The pipeline I present in this thesis empowers researchers to answer their perceptual questions by computing the scales themselves, choosing the dimensionality, and interpreting them with uncertainty in mind. They can reconsider questions previously investigated using cumbersome experimental paradigms or limited, for example one-dimensional, analysis methods and use ordinal embedding methods to view these questions from a new perspective.

Zusammenfassung

Die objektive Bestimmung der subjektiven Reizintensität beschäftigt die Wissenschaft seit mehr als 100 Jahren. Die neueste Generation dieser sogenannten psychophysischen Skalierungsverfahren kombiniert experimentelle Aufgaben, in denen Versuchspersonen die Ähnlichkeit von Reizen vergleichen, mit ordinalen Einbettungsalgorithmen, wie sie im Bereich des maschinellen Lernens entwickelt wurden, um eine robuste und mehrdimensionale Koordinatendarstellung der Reizwahrnehmung zu erhalten. Die korrekte Anwendung dieser Skalierungsverfahren erweist sich selbst für Expertinnen und Experten als schwierig, da viele Lücken in der Anwendung der Algorithmen und der Interpretation der Ergebnisse bestehen.

In dieser Dissertation beschreibe ich eine Pipeline, um psychophysische Skalierung mittels ordinaler Vergleiche und maschineller Lernverfahren zugänglich zu machen. Dazu stelle ich zunächst eine Open Source Python Toolbox vor, die die wichtigsten Algorithmen und Methoden in einfach zu bedienenden und effizienten Implementierungen zur Verfügung stellt. Dann schlage ich ein Verfahren vor, um die wichtige Wahl der Dimensionalität der Skala auf der Grundlage statistischer Überlegungen zu vereinfachen, sowie Analysemethoden, um die Stabilität einer Skala zu schätzen und wissenschaftliche Schlussfolgerungen daraus zu ziehen. Ich schließe die Arbeit mit einer neuartigen Anwendung der vergleichsbasierten Skalierungsmethoden in einem Virtual-Reality-Experiment ab. Dabei messen wir die wahrgenommene Stärke der optischen Verzerrung von Gleitsichtbrillen, die zu schwerwiegenden Nebenwirkungen wie Schwindel führen kann.

Die Pipeline, die ich in dieser Thesis vorstelle, ermöglicht es Forscherinnen und Forschern, ihre eigenen Wahrnehmungsfragen zu beantworten, indem sie selbst Skalen berechnen, die Dimensionalität auswählen und unter Berücksichtigung von Unsicherheit interpretieren. Fragen, die bisher mit aufwändigeren experimentellen Paradigmen oder eingeschränkten, z.B. eindimensionalen Analysemethoden untersucht wurden, können nun in einem neuen, umfassenderen Licht betrachtet werden.

List of publications

Publications described in this dissertation

Künstle, D.-E., & von Luxburg, U. (2024). Cblearn: Comparison-based machine learning in python. *Journal of Open Source Software*, *9*(98), 6139

Künstle, D.-E., von Luxburg, U., & Wichmann, F. A. (2022a). Estimating the perceived dimension of psychophysical stimuli using triplet accuracy and hypothesis testing. *Journal of Vision*, *22*(13), 5

Sauer, Y., Künstle, D.-E., Wichmann, F. A., & Wahl, S. (2024). An objective measurement approach to quantify the perceived distortions of spectacle lenses. *Scientific Reports*, *14*(1), 3967

Additional publications

Sering, K., Weitz, M., Shafaei-Bajestan, E., & Künstle, D.-E. (2022). Pyndl: Naïve discriminative learning in python. *Journal of Open Source Software*, *7*(80), 4515

Huber, L. S., Künstle, D.-E., & Reuter, K. (2024). Tracing truth through conceptual scaling: Mapping people's understanding of abstract concepts *[PREPRINT]*

Conference contributions

Künstle, D.-E., von Luxburg, U., & Wichmann, F. A. (2022b). Estimating the perceived dimensionality of psychophysical stimuli using a triplet accuracy and hypothesis testing procedure. *Journal of Vision*, *22*(14), 3331

Schönmann, I., Künstle, D.-E., & Wichmann, F. A. (2022). Using an odd-one-out design affects consistency, agreement and decision criteria in similarity judgement tasks involving natural images. *Journal of*

Vision, *22*(14), 3232

Künstle, D.-E., & Wichmann, F. A. (2023). Measuring lightness constancy with varying realism. *Journal of Vision*, *23*(9), 5281

Sauer, Y., Künstle, D.-E., Wichmann, F., & Wahl, S. (2023b). Psychophysical scale of optical distortions of multifocal spectacle lenses. *Journal of Vision*, *23*(9), 5215

Acknowledgments

First of all, I would like to thank my advisors, Felix Wichmann and Ulrike von Luxburg, who guided me through this thesis and whose support I could always count on. I hope to take some of Felix's unique scientific mind and Ulrike's leadership talent to become half as good a mentor someday.

My colleagues in the NIP and TML groups made it a pleasure to come to the office or go abroad. I would like to thank Uli and Silke, who can help you with any situation; Thomas, with whom every trip is a pleasure; and Kristof, Robert, Solveig, Lena, and all the others for discussions around the coffee machine. I'm grateful to all my students. Fynn, Inés, Johannes, Line, Regina, Tanja, Edward and Barbara taught me more than I taught them.

In addition to my groups, I have learned so much from my academic network. I would like to thank Benedikt Ehinger, Marianne Maertens, Guillermo Aguilar, Tom Wallis, Marc Weitz, Tino Sering, Lukas Huber, Kevin Reuter, Yannick Sauer, and Siegfried Wahl for their advice, feedback, and fruitful collaboration.

Besides many individuals, the institutionalized machine learning community in Tübingen strongly influenced this work. Many ideas were born during talks and journal clubs, or at IMPRS-IS community events. In particular, the financial support from the ML4Science cluster and the Tübingen AI Center allowed me to explore new avenues.

Apart from this exceptional academic support, ups and downs are always part of the PhD journey—and above all, my friends and family have carried me through the frustrations and celebrated the successes with me. Throughout my school and university years, my parents, Birgit and Klaus, gave me all the freedom and support I needed to explore my paths, for which I thank them from the bottom of my heart. Finally, my beloved wife Lena and my son Mattis bring true happiness into my life and give me the strength to continue on this path every day.

Contents

1 Introduction

1.1 Objective measurements of the subjective perception

One accurate measurement is worth a thousand expert opinions.
— Grace Hopper

People have been measuring their world since ancient times; there appears to be an incredible fascination with assigning numerical values to size, weight, and distance. Today, people quantify their lives thoroughly with smartphones that constantly count steps or smartwatches tracking their heart rate. In addition to these measurements of physical activity and physiology, many people like to quantify their sensations and emotions as part of personal health tracking or to share their inner world with others:

Rate your sleep quality 

Did you like the wine? 
How was the video quality of the meeting? 

The subjective component of this "feeling data" is evident, in contrast to the physical measurement of physiology, which is limited mainly by the quality of the measuring equipment. From a scientific perspective, however, even the subjective impressions ought to be objectively quantified. Psychophysical methods[1] combine strictly controlled behavioral experiments with data analyses and statistical models (Fechner, 1860). The model in our case, the *psychophysical scale*, numerically describes the perceived intensity and thus not only tells us that the Pinot Grigio tasted better than the Chardonnay but how much better.

[1] Despite the name psycho*physics*, there are also many examples in which there is no physical counterpart to the perceived intensity—aesthetics, taste, and artistic qualities can also be measured with psychophysical methods (e.g., Ribe, 2022; Schroeder, 1984; Wijntjes et al., 2020). The interested reader can find a comprehensive introduction to psychophysical methods in Wichmann and Jäkel (2018).

In fact, psychophysical scales affect our daily lives, even without us noticing them. For example, if we double the volume of our music player, the music should sound twice as loud. This scale of perceived loudness is logarithmically related to the physical sound pressure (Goldstein, 2007). Similar internal color and brightness scales calibrate the color space of computers such that images are displayed realistically (Fairchild, 2013).

On the other hand, there are many applications where psychophysical scales would be helpful but are not yet used, such that improve-

ments in psychophysical methods have an enormous scientific and economic potential for quantifying real-world multidimensional perception. For example, when a new movie is shot, the colors of the raw material must be mapped to standard metrics of display devices (e.g., cinema or TV) while preserving the artistic intent. Because computational color models cannot predict observer preferences, this mapping of perceived colors is still a manual task of human experts on "a shot-by-shot, object-by-object basis" (Zamir et al., 2021). Similarly, the quality of optical devices such as microscopes, binoculars, and eyeglasses is still determined either by physical properties or by subjective "feeling" after unsystematic use by designers and managers—even though the objective perceptual properties ultimately matter.

Psychophysical methods quantify perception through behavioral experiments and data analysis routines (Wichmann & Jäkel, 2018). Different aspects of perception can be studied. The most traditional methods investigate the discriminative ability of observers via intensity *thresholds* or just-noticeable differences (JND; Fechner, 1860) that typically indicate the minimum intensity difference at which two stimuli can be discriminated *reliably*, i.e. with 75% accuracy. This statistical perspective on discrimination is grounded in Signal Detection Theory (SDT; Green & Swets, 1966). Psychophysical scales, however, measure the *appearance* of a stimulus and the perceptual magnitude of *suprathreshold* stimulus differences, i.e., of stimuli that can be easily discriminated because their perceived dissimilarity is larger than the discrimination threshold (Gescheider, 1988; Maloney & Knoblauch, 2020). In many applications, scales are much more interesting than thresholds. Because they not only tell us that Pinot Grigio tastes different from Chardonnay but also quantify how much better it tastes, they make it possible to relate taste improvement to other parameters such as price difference.

The experimental task largely determines the scientific insights that can be gained. Recently, tasks that compare the order of stimulus similarity have become very popular (e.g., Aguilar et al., 2017; Brown et al., 2011; Charrier et al., 2007; Demiralp et al., 2014; Devinck & Knoblauch, 2012; Fleming et al., 2011; Hebart et al., 2020; Knoblauch et al., 2020; Lagunas et al., 2019; Maloney & Yang, 2003; Obein et al., 2004; Roads & Love, 2021; Roads & Mozer, 2019; Rogers et al., 2016; Sauer et al., 2024; Waraich & Victor, 2024; Wills et al., 2009). These so-called ordinal comparisons include, for example, the triad or triplet question "Is stimulus i more similar to j or k?" (Torgerson, 1952), the quadruplet comparison "Which pair is more (dis)similar" (e.g., Maloney & Yang, 2003), or the odd-one-out question (e.g., Hebart et

al., 2020). Unlike other tasks, observers do not have to learn to rate the similarity of the stimulus set consistently. Observers have reported that ordering similarities is intuitive, resulting in less training time and more consistent responses (Aguilar et al., 2017; Demiralp et al., 2014; Li et al., 2016; Roads & Mozer, 2019; Wichmann et al., 2017).

Traditionally, triplets are asked repeatedly to infer pairwise similarities based on the response rate and to compute scales from these similarities (Torgerson, 1952). Maximum Likelihood Difference Scaling requires only a selection of triplets to estimate the perceived stimulus intensities (Maloney & Yang, 2003). However, due to technical limitations, these intensities must have a monotonic progression. Haghiri et al. (2020) pointed out that flexible algorithms for determining scales with arbitrary geometry can be found in machine learning as so-called ordinal embedding methods. These algorithms can estimate even multidimensional scales from a subset of triplet responses and thus open up new paths, but also new questions, for psychophysical scaling.

1.2 *Pushing the envelope of psychophysical scaling*

Psychophysical scaling has a long tradition in psychological research. This section will look at a number of different scaling procedures to understand their shortcomings and the advantages of methodological developments. We do not proceed in strict chronological order, and we do not claim to provide an exhaustive overview. Still, we group examples of scaling methods according to the experimental task and the possible interpretation, i.e., what the scale can tell us about perception[2].

[2] "What the scale can tell us" is a fundamental question being investigated by the research field of *measurement theory* (Krantz et al., 2006; Suppes & Krantz, 2007), which specifies theorems and axiom systems. These axioms can be used in practical applications: For example, Knoblauch and Maloney (2012b, ch. 7, pp. 223-228) developed diagnostic tests for their scaling method to determine whether the desired interpretation is reasonable from the available data.

1.2.1 *Measuring perceived intensity*

In everyday applications, perception is often queried directly via scores, such as "rate the audio quality from 1 to 5". This so-called *magnitude estimation* (Stevens, 1957, 1960) and other direct scaling methods make it possible to read the perceived intensity directly from the observer's response. As an alternative to ratings, direct methods can ask to *adjust* a stimulus attribute (Merkel, 1888; Stevens, 1959) or to *partition* the attribute into equally perceived sections (Gescheider, 2013b). While these tasks seem intuitive, they are scientifically controversial: It has been shown that observer responses are biased easily by training, instructions, or stimuli selection (Beck & Shaw, 1965; Poulton, 1968; Robinson, 1976). In addition, Shepard (1981) questioned the observer's ability to report a psychological magnitude directly. Instead, it was argued that they would instead report learned labels. In addition, there are known biases in observers' responses that might be caused by the

instructions, the preceding trials, or the stimulus frequency (Gescheider, 1988).

INDIRECT SCALING methods avoid these problems by comparing only a few stimuli shown per trial to calculate the scale from all responses (Gescheider, 2013a). For example, Fechner's JND scales (Fechner, 1860) are based on the idea that the unit distance of the scale corresponds to an intensity difference that can be reliably discriminated; the concatenation of repeated JND measurements at adjacent intensities results in the scale[3]. Another method that uses discriminability to infer the perceived distance is *Thurstonian scaling* or *paired comparison scaling* (Thurstone, 1927). Thurstonian scaling models the similarity of two stimuli with a normal distribution, fitted to the frequency with which the observers could discriminate between them; the psychological unit distance is not the JND but the z-score of the normal distribution. Thurstone's approach does not require active manipulation of stimulus intensity and thus can measure abstract concepts such as aesthetics or immutable stimuli such as works of art (e.g., Wijntjes et al., 2020). However, constructing these *discrimination scales* by Fechner or Thurstone requires by design that stimulus pairs are difficult to distinguish so that the observer confuses them. This near-threshold similarity of stimuli makes the task demanding for the observer and limits the choice of stimuli. For example, scaling of clearly distinguishable intensities or discrete stimuli such as object images or words may not be possible.

[3] One of the big open questions in psychophysics is to what extent thresholds can be derived from scales and vice versa. The state of research shows mixed evidence in this case (e.g., Aguilar et al., 2017; Devinck & Knoblauch, 2012; Ross, 1997).

1.2.2 *Representing multidimensional perceptual spaces*

IT IS typically inadequate to characterize the appearance of stimuli like material samples, colored patches, or photographed things with a single value. For example, patches of textured materials like leather, orange skins, or reptile scales could be compared based on their glossiness and bumpiness. A comprehensive representation of their perceived similarity might be characterized by multiple values, so-called *dimensions* of the perceptual space. Ho et al. (2008) suggests using a conjoint measurement approach (Luce & Tukey, 1964) to determine these dimensions and their interaction, called *Maximum Likelihood Conjoined Measurement (MLCM)*. In two experiments, they ask observers to discriminate two surfaces that differ by their specularity or 3D structure. MLCM infers perceived glossiness and bumpiness as the parameters of a model, predicting the respective discriminations. Besides the simple addition of those parameters, including interactions between perceived dimensions with multiplication is straightforward. Like all discrimination scales, MLCM is relatively restrictive in terms of stim-

ulus selection—the stimuli must be similar enough to be confusable and specified by (physical) properties that can be manipulated independently.

SIMILARITY between stimuli can be used to study perception considerably more flexibly and detached from the physical properties. *Multidimensional scaling* (MDS) assigns coordinates to each stimulus in the putative perceptual space such that similarly perceived stimuli are close to each other, i.e. have a small Euclidean distance[4]. For example, Figure 1.1 illustrates hue perception like a 2D "city map" with arbitrary axes, where similar perceived colors are close together (compare pink and red with blue and green). It is important to note that this scale is an interval scale (Stevens, 1946). No natural origin or scaling exists, such that the identical similarities are represented by shifted, rotated, flipped, or scaled coordinates.

Most scaling methods discussed so far estimate perceived similarity from a non-zero error rate in discrimination or detection experiments. With MDS, however, this error rate no longer matters, allowing experiments to use stimuli with suprathreshold differences where the stimuli are clearly distinguishable. There are several behavioral experiments to obtain values for perceived stimulus similarity. Most commonly, similarity is obtained by repeatedly rating pairs of stimuli (e.g. Ekman, 1954; Shepard et al., 1975). Alternative methods include obtaining similarity from confusion rates between presented and intended stimuli (e.g., Miller & Nicely, 1955) or from a spatial arrangement of a set of stimuli (Goldstone, 1994).

Algorithmically, *Torgerson scaling* (or Principal Coordinates Analysis; Torgerson, 1952) is one of many solutions for estimating coordinates from a matrix of pairwise stimulus dissimilarities. The coordinates are derived by eigenvalue decomposition of the dissimilarity matrix. As with other direct methods, it can be assumed that similarity or dissimilarity ratings are only a distorted view of perception (cf. subsection 1.2.1). For these data, it is suggested to use non-metric MDS methods (Kruskal, 1964a, 1964b; Shepard, 1962), which transform the dissimilarities with a monotonic function. The final scale is determined by alternating between the fitting of the coordinates and the transformation.

[4] The geometric representation of perception is undoubtedly controversial. For some stimuli, perceived similarities may not satisfy the conditions for a metric. For example, the triangle inequality states that the direct distance between two points $d(A, B)$ is never longer than the distance via a third point $d(A, C) + d(C, B)$. Measuring $d(A, B)$ and $d(A, C)$ on different object properties might violate the triangle inequality (Tversky & Gati, 1982). The interested reader can find a more comprehensive review of psychological space representations in Roads and Love (2024).

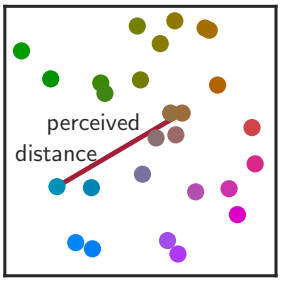


Figure 1.1: A 2D scale of hue perception (simulated). Each point represents a stimulus such that the Euclidean distance between points corresponds to the perceived dissimilarity.

1.2.3 *Comparing similarities*

INSTEAD of asking for similarity ratings, combined with the biases of direct methods described above, ordinal comparisons between pairs of stimuli are becoming increasingly popular (Aguilar et al., 2017; Hebart et al., 2020; Lagunas et al., 2019; Maloney & Yang, 2003; Roads &

Mozer, 2019; Suresh et al., 2023; Waraich & Victor, 2022; Wills et al., 2009, e.g.,). The *method of triads* or *triplet tasks*, as introduced by Torgerson (1952), asks the observer whether "stimulus i is more similar to stimulus j than to stimulus k". Triplet experiments are intuitive and thus typically preferred by trained and untrained observers over threshold approaches (Aguilar et al., 2017; Wichmann et al., 2017); scales can be estimated robustly even with a fraction of the possible triplet questions (Demiralp et al., 2014; Haghiri et al., 2020; Maloney & Yang, 2003).

Some studies use alternative ordinal comparison tasks, like the one in Figure 1.2, to make the task even more intuitive or faster than the triplets (Roads & Mozer, 2019). However, most comparative data can be converted to triplets for analysis purposes. For example, "stimulus i is the odd-one-out of i, j, k" corresponds to the triplets "j is more similar to k than i" and "k is more similar to j than i".

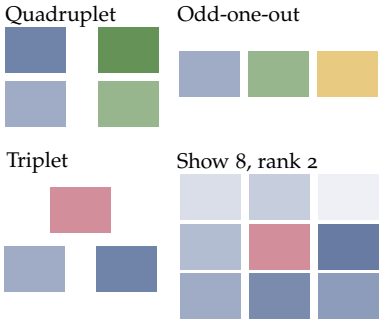


Figure 1.2: Variations on ordinal comparison tasks. The *quadruplet* task compares the similarity of stimulus pairs (blue vs. green). The *odd-one-out* task identifies the stimulus most dissimilar to all others. The *triplet* and *show 8, rank 2* tasks compare a reference (red) to the other stimuli (blue).

While Torgerson (1952) showed triplets repeatedly to estimate similarity from response ratios, modern scaling algorithms can estimate scale from a (random) sample of triplet responses. These modern algorithms optimize the coordinates to satisfy as many triplets as possible, i.e., stimulus i is represented closer to j than k.

In vision science, comparison-based scaling approaches are well established. Most notably, the *Maximum Likelihood Difference Scaling* (MLDS; Knoblauch & Maloney, 2008, 2012a; Maloney & Yang, 2003) algorithm has been used to investigate topics ranging from image quality metrics to attributes of materials or colors (e.g., Aguilar et al., 2017; Brown et al., 2011; Charrier et al., 2007; Devinck & Knoblauch, 2012; Fleming et al., 2011; Knoblauch et al., 2020; Obein et al., 2004; Rogers et al., 2016). MLDS obtains a monotonic 1D scale using linear regression on restricted triplet data, limited to trials where the perceived stimulus intensities x are ordered such that $x_i < x_j < x_k$.

Many questions in perceptual research require a multidimensional representation, however—or the ability to obtain one when needed. It is therefore not surprising that a large number of MDS-like methods have been developed that learn from similarity comparisons (Agarwal et al., 2007; Bimler et al., 2000; Hebart et al., 2020; Roads & Mozer, 2019; Takane, 1978; Waraich & Victor, 2022). These methods enable the study of materials in multidimensional space (Schmid & Anderson, 2017; Wills et al., 2009), but also the measurement of object (Demiralp et al., 2014; Hebart et al., 2020; Roads & Love, 2021) or concept spaces (Huber et al., 2024; Waraich & Victor, 2024). More recently, large-scale online studies have been conducted that collectively analyze the trials of many observers to scale thousands of object images or rendered

materials (Hebart et al., 2020; Lagunas et al., 2019; Roads & Love, 2021; Schmid & Anderson, 2017). Some of these large-scale studies use adaptive trial selection (or active learning) to minimize the number of trials (Jamieson et al., 2015; Roads & Love, 2021; S. Sievert, 2021; Tamuz et al., 2011).

1.3 *Machine learning for ordinal comparison data*

METHODS to estimate similarity purely from their ordinal comparisons are not unique to perceptual science but are an active field of machine learning research.

COMPARISON-BASED methods were developed for most machine learning applications such as metric learning (Liu et al., 2012; Shi et al., 2014), clustering (Ghoshdastidar et al., 2019; Haghiri et al., 2017; Perrot et al., 2020), classification, or regression (Haghiri et al., 2018). Contrastive learning methods estimate a mapping from features to representation based on triplets and are popular for training deep-learning-based image classifiers (Chen, Kornblith, Norouzi, & Hinton, 2020; Chen, Kornblith, Swersky, et al., 2020).

Haghiri et al. (2020) point out that a particular class of comparison-based learning methods, the *ordinal embedding* algorithms, can be used to estimate psychophysical scales from ordinal comparisons. Ordinal embedding algorithms (e.g., Agarwal et al., 2007; Tamuz et al., 2011; Terada & von Luxburg, 2014; van der Maaten & Weinberger, 2012) estimate Euclidean coordinates from triplets and quadruplets, much like the comparison-based MDS methods—for this reason, we will use embedding and (multidimensional) scale interchangeably in the following.

1.3.1 *Ordinal embedding algorithms*

ORDINAL EMBEDDING algorithms aim to find D-dimensional points $\mathbf{x}_1, \ldots, \mathbf{x}_N \in \mathbb{R}^D$ that are consistent with as many collected comparisons as possible. When encoding a triplet comparison "stimulus i is more similar to stimulus j than k" by the order of the stimulus indices (i, j, k), the representation is "consistent" if the order of distances d conforms to the reported order of similarity[5]:

$$d(\mathbf{x}_i, \mathbf{x}_j) < d(\mathbf{x}_i, \mathbf{x}_k)$$

Given a set of triplets $\mathcal{T}$, we can count the inconsistent triplets to determine the triplet error ($\mathbb{1}$ is 1 if the comparison is true, 0 otherwise).

[5] In this work, the distance d is Euclidean to simplify downstream applications on the final embedding, such as visualization or clustering.

$$\frac{1}{|\mathcal{T}|} \sum_{(i,j,k)\in\mathcal{T}} \mathbb{1}[d(\mathbf{x}_i, \mathbf{x}_j) < d(\mathbf{x}_i, \mathbf{x}_k)]$$

An alternative notation for triplet comparisons explicitly states the observer's judgment as $r = 1$ (consistent with the triplet) or $r = 0$. This way, we can hypothetically reformulate the ordinal embedding problem as a binary classification problem, predicting r from (i, j, k). The embedding—or "model parameter"—x can be found by numerically minimizing the classification loss $\mathcal{L}$.

$$\begin{aligned} \hat{r} &:= \mathbb{1}[d(\mathbf{x}_i, \mathbf{x}_j) < d(\mathbf{x}_i, \mathbf{x}_k)] \\ &\underset{\mathbf{x}_1 \ldots \mathbf{x}_N \in \mathbb{R}^D}{\arg\min} \sum_{(i,j,k),r\in T} \mathcal{L}(r, \hat{r}) \end{aligned}$$

In practice, this discrete loss function cannot be numerically optimized efficiently. Instead, the discrete loss is replaced by convex relaxations (Agarwal et al., 2007; Terada & von Luxburg, 2014) or smooth, probabilistically motivated loss functions (Tamuz et al., 2011; van der Maaten & Weinberger, 2012). Jain et al. (2016) showed formally that, in general, convex optimization can estimate ordinal embeddings.

The loss functions of ordinal embedding algorithms encode some assumptions about the response distribution. For example, these are the loss functions (sometimes called the *noise model*) of Crowd Kernel Learning (*CKL*, Tamuz et al., 2011), Stochastic Triplet Embedding (*STE*, van der Maaten & Weinberger, 2012), and Soft Ordinal Embedding (*SOE*, Terada & von Luxburg, 2014). A comprehensive collection of algorithm descriptions can be found in Vankadara et al. (2021).

$$\begin{aligned} CKL: &\quad \frac{d(\mathbf{x}_i, \mathbf{x}_j)^2 + \mu}{d(\mathbf{x}_i, \mathbf{x}_j)^2 + d(\mathbf{x}_i, \mathbf{x}_k)^2 + 2\mu} \\ STE: &\quad \frac{e^{-d(\mathbf{x}_i, \mathbf{x}_j)^2}}{e^{-d(\mathbf{x}_i, \mathbf{x}_j)^2} + e^{-d(\mathbf{x}_i, \mathbf{x}_k)^2}} \\ SOE: &\quad \max\{0, (d(\mathbf{x}_i, \mathbf{x}_j) - d(\mathbf{x}_i, \mathbf{x}_k) + \mu)^2\} \end{aligned}$$

In psychophysical scaling, we usually also make assumptions about the observer's noise distribution in perception and response. For example, a common assumption is that errors occur when both stimulus pairs are (almost) equally similar (cf. the noise model in the MLDS literature, Maloney & Yang, 2003). This assumption fits all loss functions but via the difference in distances in SOE and the distance ratio in STE/CKL. For some stimuli, however, it is a common assumption that the variation increases with the stimulus intensity (e.g., Poisson

noise; Egan, 1975; Kaernbach, 1991), which corresponds to the noise model of CKL and STE.

1.3.2 *Theoretical insights*

In addition to new algorithms, the machine learning literature contributes to the theoretical investigation of comparison-based learning methods. These insights can be transferred into practice to make statements about the reliability or the correct application of ordinal embedding methods.

It is not trivial that ordinal comparisons are sufficient to reconstruct an embedding. Shepard (1964) already observed that with an increasing number of points (and thus, comparisons), a ranking of the distances sufficiently narrows down the "wiggle" room to such an extent that they fix the points. With triplets, we can intuitively think of the "wiggle" room as follows: Every triplet bisects the embedding space R^D, such that one side is closer to $\mathbf{x}_j$, shown in red in Figure 1.3 and contains the target $\mathbf{x}_i$ (Jamieson & Nowak, 2011). Further triplets will continue to trim the putative space of $\mathbf{x}_i$ until its point is determined.

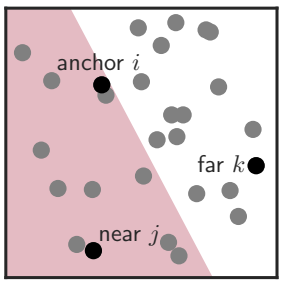


Figure 1.3: A triplet is bisecting the space of a 2D embedding. The red area shows the half-space closer to the "near" point, containing the "anchor".

Formal proofs of this intuition were provided by recent works. In the limit, ordinal embedding algorithms can reconstruct a *unique* set of points from comparisons alone up to the similarity transformations, i.e., translation, rotation, scaling, and mirroring of the embedding (Arias-Castro, 2017; Kleindessner & von Luxburg, 2014)

However, the number of comparisons, and thus the number of trials in the experiment, required for this reconstruction is also crucial for an application in psychophysical scaling. If an embedding of N stimuli would require all $3\binom{N}{3}$ triplets, there are half a million possible trials for 100 stimuli. Jamieson and Nowak (2011) have observed and conjectured that the required number of triplets for D-dimensional embeddings is much smaller and does not grow any faster than $\mathcal{O}(DN \log N)$, visualized in Figure 1.4. Jain et al. (2016) formally proved that this growth rate indeed bounds the prediction error of the embedding for noisy triplets. Therefore, ordinal embedding experiments typically require far fewer trials than MDS experiments since only a fraction of all possible pairs or triplets of stimuli have to be queried.

Additional theoretical works investigated the triplet sampling. For example, Terada and von Luxburg (2014) could show that local information, i.e., nearest neighbor comparisons, are sufficient to recover the embedding globally. Careful sampling schemes can decrease the total number of trials: Landmark-based approaches require $\mathcal{O}(DN \log M)$ ($M \ll N$) triplets to approximate the embedding (Anderton & Aslam, 2019). Alternatively, active sampling algorithms collect only the most

informative triplets (e.g., Tamuz et al., 2011).

1.4 *A pipeline for comparison-based scaling*

Overall, ordinal embedding algorithms are a powerful tool for estimating psychophysical scales from ordinal comparisons, which can estimate multidimensional representations even from random subsets of the possible comparisons. However, with their great flexibility comes insecurity about how to use them correctly. Whether it is the choice of implementation, dimensionality, validation method, or experimental design, there are many open questions, little evidence, and no guidance—and thus the risk for practitioners to make mistakes. Haghiri et al. (2020) conclude their work with recommendations and open questions about the practical application of ordinal embedding methods in psychophysical scaling, which I take up, develop further, or even correct in this thesis.

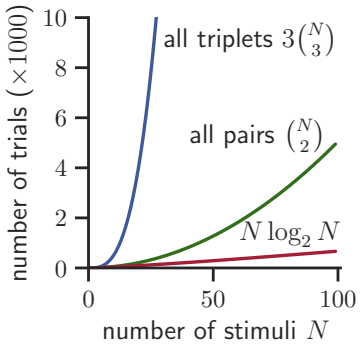


Figure 1.4: The number of triplets and pairs grows fast with the number of stimuli N. However, a subsample of $O(DN \log N)$ triplets is sufficient to estimate the embedding (Jain et al., 2016).

Based on the theoretical results on error bounds described above, they suggest choosing the number of stimuli according to $DN \log N$. The only implementation they can recommend for estimating the scale is an undocumented MATLAB script collection due to a lack of alternatives. Since this makes it difficult for new users to get started with the methodology, I am correcting the recommendation in this paper by presenting a comprehensively documented and tested toolbox. The recommendation of Haghiri et al. (2020) for determining the dimensionality and quality of the scale is based on observing the cross-validated triplet error. This recommendation is also examined in more detail in the following chapters and is partly extended and partly corrected. The open issues regarding the application of ordinal embedding methods refer primarily to the estimation of confidence intervals, approaches to interpretation, and conjoint measurement. I address the former in a separate chapter of this thesis, while the others are covered in the discussion.

Since such recommendations for the use of ordinal embedding methods have been available for four years, why do many researchers prefer other methods like MLDS (Maloney & Yang, 2003), which limits the stimulus selection (vary only one property), the experimental task (only certain triplets and quadruplets), and especially the analysis (only 1D scales)?

What MLDS offers is not just an algorithm, but a scaling pipeline consisting of an open source R implementation (Knoblauch & Maloney, 2008) and various procedures for data and model analysis (Knoblauch & Maloney, 2012a), comprehended by numerous successful research

applications (e.g., Charrier et al., 2007; Devinck & Knoblauch, 2012; Fleming et al., 2011; Knoblauch et al., 2020; Obein et al., 2004; Rogers et al., 2016). This comprehensive approach of software, methods, and applications that complement each other makes it easier for new users to get started and gain confidence. The machine learning community might consider this approach of lowering the entry threshold as "democratization of AI use" (Ahmed et al., 2020; Seger et al., 2023). The AI democratization movement aims to make AI tools more accessible and usable. For AI in scientific applications, this could be achieved through free and easy-to-use software (Seger et al., 2023; Sundberg & Holmström, 2023), automated algorithm configuration (AutoML; Hutter et al., 2019; Imrie et al., 2023), or improved interpretability and explainability of the model (Ahmed et al., 2020; Roscher et al., 2020). For psychophysical scaling with ordinal embedding methods, I am convinced that democratization can be created through a pipeline of software and methods along with applications.

This thesis describes a pipeline for psychological scaling using ordinal comparisons with ordinal embedding algorithms. This pipeline consists of four parts: a software toolbox, methods for estimating the scaling dimension and its stability, and the application of procedures in a new domain of psychophysical scaling.

In chapter 2 I present the `cblearn` toolbox, which implements the most relevant ordinal embedding algorithms and utilities for generating, loading, and transforming ordinal comparison datasets[6]. Through code examples and a simple API that follows conventions of the Python machine learning ecosystem, I aim to significantly lower the barrier to testing the methods. The most critical configuration of ordinal embedding algorithms is the choice of the dimensionality of the embedding, for which I present our data-driven procedure in chapter 3[7]. The embedding obtained from the algorithm can be misinterpreted easily. To ensure that it can be trusted, i.e., that it does not contain random patterns but only those from the data, in chapter 4, we consider the stability of ordinal embeddings and present approaches to quantify it. Finally, in chapter 5, I demonstrate the robustness and flexibility of the method by presenting our study on scaling perceived lens distortions based on sequentially presented triplets in a Virtual Reality environment[8]. All results, including their limitations and potential, are discussed in chapter 6. I conclude the discussion with potential applications of this work beyond psychophysics and the necessary further development of this machine learning framework for psychophysical scaling.

[6] Künstle, D.-E., & von Luxburg, U. (2024). Cblearn: Comparison-based machine learning in python. *Journal of Open Source Software*, *9*(98), 6139

[7] Künstle, D.-E., von Luxburg, U., & Wichmann, F. A. (2022a). Estimating the perceived dimension of psychophysical stimuli using triplet accuracy and hypothesis testing. *Journal of Vision*, *22*(13), 5

[8] Sauer, Y., Künstle, D.-E., Wichmann, F. A., & Wahl, S. (2024). An objective measurement approach to quantify the perceived distortions of spectacle lenses. *Scientific Reports*, *14*(1), 3967

2 *Toolbox for comparison-based learning*

The loftier the building, the deeper must the foundation be laid.
— Thomas à Kempis

When I first started working on ordinal embedding methods, we found implementations of ordinal embedding methods primarily as collections of scripts in different programming languages.

This, of course, makes it difficult to compare and improve methods. Even worse, it makes it difficult to get started with ordinal embedding and to apply it to the analysis of psychophysical data. That's why I started to develop a Python library that provides a standardized representation of comparison data installed with "batteries included", i.e. access to data sets and implementations of the most important algorithms.

We have described this software package in a published article[1], licensed under CC BY 4.0, the form and content of which is largely reproduced in this chapter. In addition, I have integrated the supplementary material of this article as new sections of this chapter and added a concluding paragraph at the end.

[1] Künstle, D.-E., & von Luxburg, U. (2024). Cblearn: Comparison-based machine learning in python. *Journal of Open Source Software*, *9*(98), 6139

Both the package and the paper are largely my work. I did the design, implementation, documentation, execution, and analysis of the simulation studies, as well as the majority of writing. Ulrike von Luxburg contributed to the scientific idea, data interpretation, and writing of the paper.

In addition, several former research group members advised me on the conceptual design, and several researchers made smaller code contributions to the library. Details about these code contributors can be found in the article acknowledgments and the GitHub contributors page.

2.1 *Introduction*

Comparison-based machine learning algorithms are used when only comparisons of similarity between data points are available but no explicit similarity scores or features. For example, humans struggle to assign *numeric* similarities to apples, pears, and bananas. Still, they can easily *compare* the similarity of pears and apples with the similarity of

apples and bananas—pears and apples usually appear more similar. There exist comparison-based algorithms for most machine learning tasks, like clustering, regression, or classification (e.g., Balcan et al., 2016; Heikinheimo & Ukkonen, 2013; Perrot et al., 2020); The most frequently applied algorithms, however, are the so-called ordinal embedding algorithms (e.g., Agarwal et al., 2007; Amid & Ukkonen, 2015; Anderton & Aslam, 2019; Ghosh et al., 2019; Tamuz et al., 2011; Terada & von Luxburg, 2014; van der Maaten & Weinberger, 2012). Ordinal embedding algorithms estimate a metric representation, such that the distances between embedded objects reflect the similarity comparisons. These embedding algorithms have recently come into fashion in psychology and cognitive science to quantify the perceived similarity of various stimuli objectively (e.g., Haghiri et al., 2020; Roads & Mozer, 2019; Wills et al., 2009).

2.2 *The toolbox*

This section presents `cblearn`, an open-source Python package for comparison-based learning (Figure 2.1). Unlike related packages, `cblearn` goes beyond specific algorithm implementations to provide an ecosystem for comparison-based data with access to several real-world datasets and a collection of algorithm implementations. `cblearn` is fast and user-friendly for applications but flexible for research on new algorithms and methods. The package integrates well into the scientific Python ecosystem; for example, third-party functions for cross-validation or hyperparameter tuning of `scikit-learn` estimators can typically be used with `cblearn` estimators. Although our package is relatively new, it has already been used for algorithm development (Mandal et al., 2023) and data analysis in several studies (Huber et al., 2024; Künstle et al., 2022a; Sauer et al., 2024; Schönmann et al., 2022; van Assen & Pont, 2022; Zhao et al., 2023).

Figure 2.1: Toolbox logo.

We designed `cblearn` as a modular package with functions for processing and converting the comparison data in all its varieties (`cblearn.preprocessing`, `cblearn.utils`, `cblearn.metrics`), routines to generate artificial or load real-world datasets (`cblearn.datasets`), and algorithms for ordinal embedding and clustering (`cblearn.embedding`, `cblearn.cluster`).

2.2.1 *Various data formats supported*

The atomic datum in comparison-based learning is the quadruplet, a comparison of the similarity δ between two pairs (i, j) and (k, l), for example, asserting that $\delta(i, j) < \delta(k, l)$. Another popular compar-

ison query, the triplet, can be reduced to a quadruplet with $i == l$. Comparison-based learning algorithms estimate classes, clusters, or metrics to fulfill as many quadruplets as possible. In ordinal embedding, for example, the problem is to find $x_i, x_j, x_k, x_l \in \mathbb{R}^d$ s.t. $\|x_i - x_j\|_2 < \|x_k - x_l\|_2 \Leftrightarrow \delta(i,j) < \delta(k,l)$.

Besides triplets and quadruplets, there are many ways to ask for comparisons. Some tasks ask for the "odd-one-out", the "most-central" object, or the two most similar objects to a reference. `cblearn` can load these different queries and convert them to triplets, ready for subsequent embedding or clustering tasks.

DIFFERENT DATA TYPES can store triplets and `cblearn` converts them internally. A 2D array with three columns for the object indices (i, j, k) stores a triplet per row. In some applications, it is comfortable to separate the comparison "question" and "response", which leads to additional response labels that are 1, if $\delta(i,j) \leq \delta(i,k)$, and -1, if $\delta(i,j) > \delta(i,k)$. An alternative format stores triplets as a 3-dimensional sparse array. These sparse arrays convert fast back and forth to dense 2D arrays while providing an intuitive comparison representation via multidimensional indexing. For example, the identical triplet can be represented as `[[i, j, k]]`, `([[i, k, j]], [-1])` or `sparse_arr[i, j, k] == 1`.

2.2.2 *Interfaces to diverse datasets*

THERE IS no Iris, CIFAR, or ImageNet in comparison-based learning—the community lacks accessible real-world datasets to evaluate new algorithms. `cblearn` provides access to various real-world datasets, summarized in Figure 2.2, with functions to download and load the comparisons. These datasets—typically comparisons between images or words—consist of human responses. Additionally, our package provides preprocessing functions to convert different comparisons to triplets or quadruplets, which many algorithms expect.

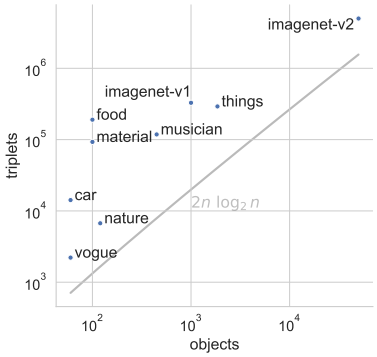


Figure 2.2: Real-world datasets that can be accessed with `cblearn` cover a wide range of sizes. Please find a detailed description and references to the dataset authors in our package documentation.

2.2.3 *Algorithms implemented for CPU and GPU*

IN THE VERSION `0.3.0`, `cblearn` implements an extensive palette of ordinal embedding algorithms and a clustering algorithm (Table 2.1); additional algorithms can be contributed easily to the modular design. Most algorithm implementations are built with the scientific ecosystem around `scipy` (Harris et al., 2020; Virtanen et al., 2020) to be fast and lightweight. Inspired by the work of Vankadara et al. (2021) , we added GPU implementations with `torch` (Ansel et al., 2024) that use automatic differentiation and stochastic optimization routines known from deep learning methods. These GPU implementations can be used

with large datasets and rapidly adapted thanks to automated differentiation.

2.2.4 *User-friendly and compatible API*

One of Python's greatest strengths is the scientific ecosystem into which `cblearn` integrates. Our package does not only make use of this ecosystem internally but adopts their API conventions—every user of `scikit-learn` (Buitinck et al., 2013; Pedregosa et al., 2011) is already familiar with the API of `cblearn`: Estimator objects use the well-known `scikit-learn` methods `.fit(X, y)`,`.transform(X)`, and `.predict(X)`. This convention provides interfaces between `cblearn`'s estimator and machine learning routines from the `scikit-learn` ecosystem. The code snippet in Figure 2.3 demonstrates this interface: The script fetches a real-world dataset, preprocesses the comparisons, evaluates the fit of an embedding model with cross-validation, and then plots an embedding estimate. Additional examples are available in the package's documentation.

2.3 *Related work*

Most comparison-based learning algorithms were implemented independently as part of a research paper (e.g., Ghoshdastidar et al., 2019; Hebart et al., 2020; Roads & Mozer, 2019; van der Maaten & Weinberger, 2012); Just a few of these implementations, for example `loe` (Terada & von Luxburg, 2014) or `psiz` (Roads & Mozer, 2019) , come in the form of software packages.

Related packages with collections of comparison-based learning algorithms focus on metric learning (cf. the `metric-learn` package with a high compatibility to scikit-learn; de Vazelhes et al., 2020) and crowd-sourced data collection, using active ordinal embedding algorithms (e.g., `NEXT` or `salmon` Jamieson et al., 2015; Sievert et al., 2023).

Table 2.1: Algorithm implementations in `cblearn`. Most of these come in multiple variants: Different backends for small datasets on CPU and large datasets on GPU as well as variations of objective functions.

Algorithm	Reference
Crowd Kernel Learning	Tamuz et al. (2011)
Fast Ordinal Triplet Embedding	Jain et al. (2016)
Generalized Non-metric MDS	Agarwal et al. (2007)
Maximum-likelihood Difference Scaling	Maloney and Yang (2003)
Soft Ordinal Embedding	Terada and von Luxburg (2014)
Ordinal Embedding Neural Network	Vankadara et al. (2021)
Stochastic Triplet Embedding	van der Maaten and Weinberger (2012)
ComparisonHC (clustering)	Perrot et al. (2020)

```
from cblearn import datasets, preprocessing, embedding
from sklearn.model_selection import cross_val_score
import seaborn as sns; sns.set_theme("poster", "whitegrid")

cars = datasets.fetch_car_similarity()
triplets = preprocessing.triplets_from_mostcentral(cars.triplet, cars.response)
accuracy = cross_val_score(embedding.SOE(n_components=2), triplets, cv=5).mean()
embedding = embedding.SOE(n_components=2).fit_transform(triplets)
fg = sns.relplot(x=embedding[:, 0], y=embedding[:, 1],
        hue=cars.class_name[cars.class_id])
```

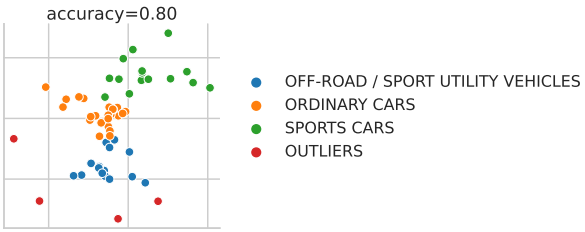


Figure 2.3: Code example using data loading, preprocessing and embedding methods from `cblearn` in interaction with other machine learning and plotting libraries.

Our package `cblearn`, on the other hand, focuses on providing comparison data and interoperable estimator implementations of the remaining areas of comparison-based learning.

2.4 *Empirical evaluation*

In this section, we want to put the goals of fast and reliable implementations to the test. We evaluate the accuracy and runtime of various ordinal embedding algorithm algorithms in `cblearn` and compare them with reference implementations by the algorithm developers, which have been recommended in the literature for practical applications (Haghiri et al., 2020, cf.). A more comprehensive evaluation of various ordinal embedding algorithms per se focusing on large data sets can be found in (Vankadara et al., 2021).

2.4.1 *Methods*

We generated embeddings of comparison-based datasets to measure runtime and triplet error as a small empirical evaluation of our ordinal embedding implementations. We compared various CPU and GPU implementations in `cblearn` with third-party implementations in *R* (Terada & von Luxburg, 2014), and *MATLAB* (van der Maaten & Weinberger, 2012). In contrast to synthetic benchmarks (e.g., compare Vankadara et al., 2021), we used the real-world datasets that can be

accessed through `cblearn`, converted to triplets. The embeddings were arbitrarily chosen to be 2D. Every algorithm runs once per dataset on a compute node (8 CPU cores; 96GB RAM; NVIDIA RTX 2080ti) with a run-time limit of 24 hours. Some runs did fail by exceeding those constraints: Our FORTE implementation was canceled by an out-of-memory error on the `imagenet-v2` dataset. The *MATLAB* implementation of tSTE timed out on `things` and `imagenet-v2` datasets. The *R* implementation of SOE on the `imagenet-v2` dataset by an "unsupported long vector" error, caused by the large size of the requested embedding.

The benchmarking scripts and results are publicly available in a separate repository[2].

[2] https://github.com/cblearn/cblearn-benchmark

2.4.2 *Is there a "best" estimator?*

Comparing the ordinal embedding estimators in `cblearn`, SOE, CKL, GNMDS, and tSTE were performing about equally well in both runtime and accuracy (Figure 2.4). The GPU implementations are slower on the tested datasets and for SOE and GNMDS noticeably less accurate.

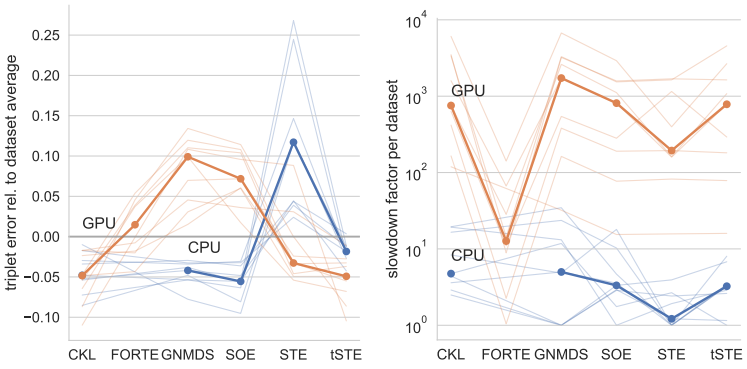


Figure 2.4: The triplet error and runtime per estimator and dataset, relative to the mean error or the fastest run. Thin lines show runs on the different datasets; the thick lines indicate the respective median. Except for STE, all CPU algorithms are able to embed the triplets similarly well. There are just minor differences in the runtime of the CPU implementations. The GPU implementations are usually significantly slower on the data sets used.

2.4.3 *When should GPU implementations be preferred?*

In terms of accuracy and runtime, our GPU implementations using the `torch` backend could not outperform the CPU pendants using the `scipy` backend on the tested datasets. However, Figure 2.4 shows the GPU runtime grows slower with the number of triplets, such that they potentially outperform CPU implementations with large datasets of 10^7 triplets and more. In some cases, the `torch` implementations show the overall best accuracy.

There are various explanations for the speed disadvantage of our `torch`-based implementations. On the one hand, it may be due to the

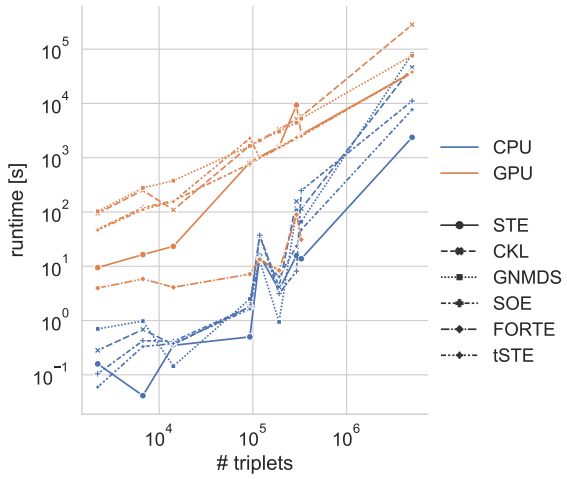


Figure 2.5: The runtime increases almost linearly with the number of triplets. However, GPU implementations have a flatter slope and thus can compensate for the initial time overhead on large datasets.

overhead of converting between `numpy` and `torch` and calculating the gradient (AutoGrad). On the other hand, it can also be due to the optimizer or the selected hyperparameters. To get a first impression of these factors, we have built minimal examples of the CKL algorithm (Tamuz et al., 2011) and estimated 2D embeddings of the Vogue Cover dataset (Heikinheimo & Ukkonen, 2013).

Figure 2.6 shows the runtimes and triplet accuracies on a standard laptop. The small markers show runs with different initialization and the bold markers the respective median performance. The CKL implementation of `cblearn` is slightly slower than the minimal version, probably due to data validation and conversion overheads. If the gradient is not provided directly but calculated automatically with `torch`'s AutoGrad functions, the minimal example, run multiple times slower. The most severe impact has to change the optimization algorithm to stochastic optimization (*Adam*, lr=10). However, following the results in previous sections, it can be assumed that this overhead is compensated for by increasing the dataset size.

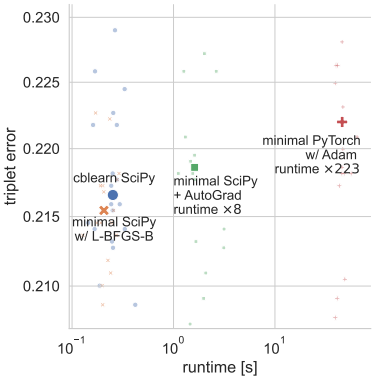


Figure 2.6: The runtime and error for different optimization methods in minimal CKL implementations. `cblearn`'s CKL implementation is shown for reference.

AN ADDITIONAL CHALLENGE of stochastic optimizers like *Adam* (Kingma & Ba, 2015) is their sensitivity to hyperparameter choices. This sensitivity is demonstrated in Figure 2.7, where the learning rate of Adam is varied for the toy example. Likewise, the performance of `torch` ordinal embedding implementations could be improved by using more sophisticated tuning of optimizer parameters.

BESIDES ALL DISCUSSIONS about runtime and accuracy, the `torch` the backend provides benefits for the maintenance and extension of the library. It uses PyTorch's automatic differentiation (Paszke et al., 2019), so that the loss gradient does not have to be explicitly defined and new

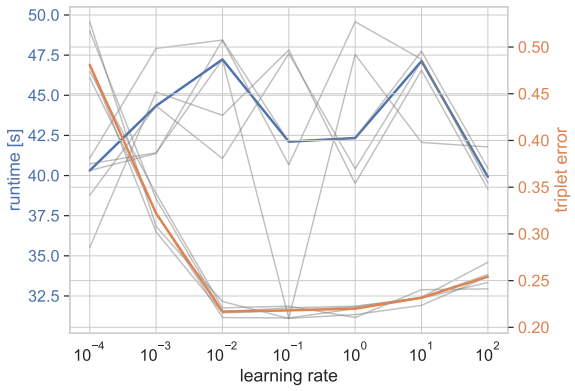


Figure 2.7: The runtime and error for different learning rates of the Adam optimizer in a minimal example with CKL estimating a 2D embedding of 60 objects.

algorithms can be implemented very quickly.

2.4.4 *How does `cblearn` compare to other implementations?*

In a small comparison, our implementations run multiple times faster with approximately the same accuracy as reference implementations (Figure 2.8). We compared our CPU implementations of SOE the corresponding reference implementations in R , `loe` (Terada & von Luxburg, 2014), and our implementation of CKL, GNMDS, STE, tSTE with the *MATLAB* of van der Maaten and Weinberger (2012). This comparison is not exhaustive, but it shows that our implementations are competitive with the reference implementations in terms of accuracy and runtime. Of course, we cannot separate the factors of algorithm implementation and runtime environment.

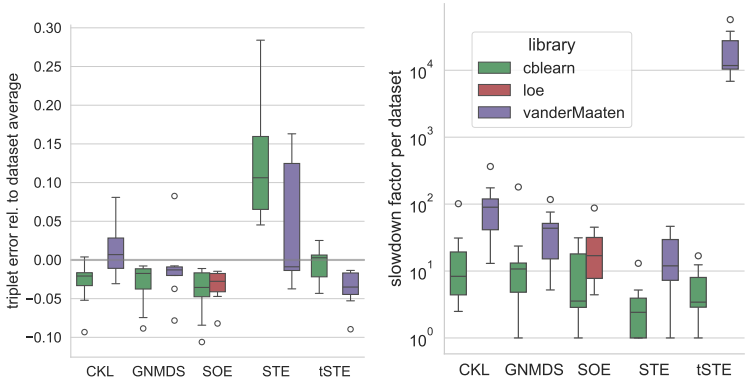


Figure 2.8: The triplet error and runtime per estimator and dataset, relative to the mean error and the fastest run. Thin lines show runs on the different datasets; the thick lines indicate the respective median. The triplet error is approximately similar for all implementations but STE. For all algorithms, 'cblearn' provides the fastest implementation.

2.5 *Conclusion*

In the previous sections, we introduced `cblearn`, a Python library that simplifies working with ordinal similarity comparison data by

providing rich interfaces to data sets, their processing, and modeling. We were able to show in an example that `cblearn`'s implementations of ordinal embedding algorithms integrate into Python's machine learning ecosystem, and in evaluations that the various implementations can match and outperform previous libraries in both accuracy and runtime.

DUE TO its modular structure and flexibility, this library enables computer scientists to explore new comparison-based methods and test them on representative datasets. Users, for example, in the field of psychophysics can easily evaluate different embedding methods using `scikit-learn`'s methods for hyperparameter search or model comparison.

This double use of the package comes with two main challenges. First, the terminology might differ in computer science and applications. For example, the term *embedding* might be better known as *(multi-dimensional) scale* in psychophysics. Second, while algorithm development might profit from low-level API, practitioners often prefer high-level functions for common tasks. In my opinion, future development of `cblearn` should extend the API to multiple levels, add application-specific modules, and extend the documentation further by adding examples and how-to guides for common tasks.

OVERALL, `cblearn` successfully contributes to research on and with comparison-based machine learning methods and has the potential to accelerate these further—the methods and data analyses in the following chapters would not have been possible in this form without `cblearn`.

3 *Dimensionality testing procedure*

Quite simply, perception is about having kids, not seeing truth.
— Donald D. Hoffman
(D. D. Hoffman et al., 2015)

A great value of ordinal embedding algorithms is that scales can be estimated in one or multiple dimensions. However, this also raises the essential question of how this dimensionality should be chosen in practice. In this chapter, I present an iterative procedure for estimating the dimensionality for typically low-dimensional psychophysical scales so that the observer responses are well represented by the embedding.

Both the content and form of this chapter equal our published article[1], licensed under CC BY 4.0. This manuscript is largely my work; I generated and analyzed the data and wrote most of the article. All article authors contributed to the scientific idea, the data interpretation, and the paper writing.

[1] Künstle, D.-E., von Luxburg, U., & Wichmann, F. A. (2022a). Estimating the perceived dimension of psychophysical stimuli using triplet accuracy and hypothesis testing. *Journal of Vision*, *22*(13), 5

Abstract

Vision researchers are interested in mapping complex physical stimuli to perceptual dimensions. Such a mapping can be constructed using multi-dimensional psychophysical scaling or ordinal embedding methods. Both methods infer coordinates that agree as much as possible with the observer's judgments so that perceived similarity corresponds to the distance in the inferred space. However, one fundamental problem of all methods that construct scalings in multiple dimensions is that the inferred representation can only reflect perception if the scale has the correct dimension. Here we propose a statistical procedure to overcome this limitation. The critical elements of our procedure are (i) measuring the scale's quality by the number of correctly predicted triplets and (ii) performing a statistical test to assess if adding another dimension to the scale improves triplet accuracy significantly. We validate our procedure through extensive simulations. In addition, we study the properties and limitations of our procedure using "real" data from various behavioral datasets from psychophysical experiments. We conclude that our procedure can reliably identify (a lower bound on) the number of perceptual dimensions for a given dataset.

3.1 Introduction

Some things feel more similar than others: Violet is bluish and reddish but not greenish, trumpet and trombone do not sound the same but are different from the violin, and platinum appears more similar to silver than gold.

One popular idea is that perceived similarities—for example, similar colors, sounds from musical instruments, or materials—correspond to distances in a coordinate system in the perceiver's mind. Methods allowing to infer the distances and the dimensionality of the internal perceptual space may thus be helpful for scientists attempting to understand perception.

STUDIES about lightness perception find, for example, that the corresponding perceptual space is not necessarily one-dimensional (Schmid & Anderson, 2017; Umbach, 2014); human observers are able to disentangle dimensions for the surface color and the illumination (Logvinenko & Maloney, 2006). Material perception research suggests that the perceptual space of materials may be spanned by subjective material properties like softness, viscosity, reflectance, or translucency (Fleming, 2017). Perceived gloss has been found to not only depend on the (physical) specular reflectance of the material but to also increase with the bumpiness of the surface (Ho et al., 2008; Kim et al., 2011; Marlow et al., 2011, 2012). Recently attempts have even been made to estimate the dimensions and the overall (high) dimensionality of object perception from large-scale crowd-sourcing studies (Hebart et al., 2020; Love & Roads, 2021; Roads & Love, 2021). In all these examples the dimensions of perceptual experience are of interest—the psychophysical scale, as it is classically and frequently also referred to (see Gescheider, 1988, 2013a, 2013b, for an overview).

THE OLDEST and most frequently used scaling algorithm for more than one dimension is (non-metric) *multi-dimensional scaling* (MDS; Kruskal, 1964a, 1964b; Shepard, 1962), which was recently accompanied by *ordinal embedding* methods from machine learning (Haghiri et al., 2020; Roads & Mozer, 2019). In contrast to other scaling approaches, such as the popular maximum-likelihood difference scaling (MLDS; Knoblauch & Maloney, 2012b), MDS and ordinal embedding can estimate multiple perceived dimensions. An approach related to MLDS but for multiple dimensions, maximum-likelihood conjoint measurement (MLCM; Ho et al., 2008), tries to obtain interpretable dimensions by additional assumptions (e.g. monotonicity and independence of perceived dimensions; Radonjić et al., 2019), whereas MDS and ordinal embedding are more exploratory in trying to find the scale that best fits the data. MDS estimates the scale by minimizing a *stress* term, measuring the agreement between the distances in the scale and dissimilarity ratings collected for (all) stimulus pairs in a psychophysical experiment.

In contrast to MDS, ordinal embedding methods use triplet comparison judgments of the form "is stimulus A more similar to B or C?"

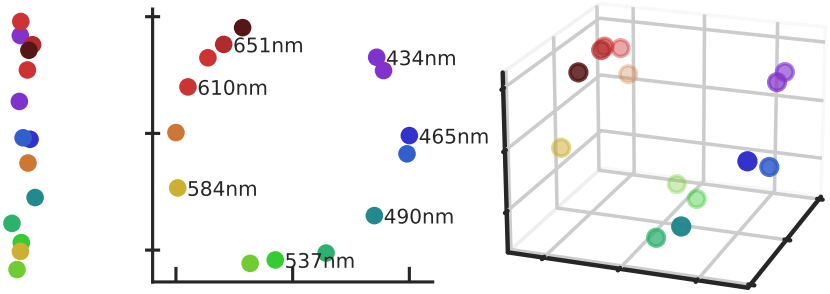


Figure 3.1: The same perceived hue similarities are represented in a one-, a two-, and a three-dimensional representation. The two-dimensional scale accurately represents the similarities in a circular structure. However, the one-dimensional scale violates some obvious similarities (e.g. orange is distant to red); the three-dimensional scale is too complex as the vertical offsets carry little to no perceptual information. The distances are based on similarity ratings between stimuli of different wavelengths (Ekman, 1954), the colors used for illustration are RGB approximations and differ from the original stimuli.

to estimate the scale (Haghiri et al., 2020). This triplet judgment task is often more intuitive for observers than other scaling tasks (Aguilar et al., 2017). Furthermore, it has to be performed on just a fraction of all possible comparisons, making ordinal embedding methods feasible with more stimuli than MDS (Haghiri et al., 2020).

However, one fundamental problem of psychophysical scaling in multiple dimensions is choosing the "correct" number of dimensions because both scaling methods, MDS and ordinal embedding, require the user to specify the dimensionality as a method parameter. Unfortunately, the scale's representation can only reflect perception if it has the "correct" dimensionality. The problem is illustrated in Figure 3.1 using perceived color similarities. The very well-known color circle is obtained only in a two-dimensional embedding. The distances in the one-dimensional embedding are distorted, so conclusions about the perceived similarities are misleading. While the distances are correct in the three-dimensional embedding, the additional third dimension carries no perceptually valid information and is thus also misleading.

Hence dimensionality is crucial in multi-dimensional perceptual scaling. The standard approach to choosing the "correct" dimensionality for MDS is visualizing the stress for different dimensions—ideally, stress should decrease and show a knee at the intrinsic (correct) dimensionality (Borg & Groenen, 2005). Perceptual studies with ordinal embedding algorithms used various approaches to determine the correct dimensionality. Some studies first estimated a high-dimensional scale with an ordinal embedding algorithm and subsequently dropped dimensions until a measure of explained variance fell below a predefined threshold (Toscani et al., 2020). Others selected the dimension where the probability of hold-out judgments, i.e. judgments which were not used to estimate the scale, is maximal (Roads & Love, 2021) or where the judgement accuracy is beyond an (arbitrary) threshold (Haghiri et al., 2020).

HOWEVER, none of the above attempts to infer the appropriate or correct dimensionality is entirely satisfactory. There is a highly subjective component in inspecting a stress graph or choosing the variance or accuracy threshold. In addition, these methods do not explicitly consider the intrinsic stochasticity of the perceptual judgments (random sampling, human factor) and the stochasticity of scaling algorithms themselves (random initialization). Thus an appropriate dimension estimation procedure should include the distribution or variation of the scale's accuracy and provide an interpretable decision criterion—a typical application of a statistical test. Such tests have already been applied to scaling models; for example, Radonjić et al. (2019) use a t-test on the cross-validated model fit of an MLCM-like scaling model to decide between the Euclidean or City-Block distance metrics.

HERE we propose a statistical procedure inspired by model selection to choose the dimensionality: Tuning the dimensionality can prevent under- and overfitting. Too simple models do not fit the data well enough; conversely, too complex models can typically fit the data but are prone to overfitting, i.e. fitting noise instead of behavior. This view transforms the dimensionality choosing problem into a model selection problem—and allows us to benefit from the extensive and time-proven model selection literature and methods. The critical elements of our suggested dimensionality estimation procedure are, first, measuring the scale's quality by the number of correctly predicted triplets (cross-validated triplet accuracy). Second, performing a statistical test to assess if adding another dimension improves triplet accuracy significantly. In order to validate this procedure, we simulated noisy and sparse judgments and assessed reliability in identifying the ground-truth dimensionality. Furthermore, we studied the properties and limitations of our procedure using "real" data from various behavioral datasets from psychophysical experiments.

We conclude that our procedure is a robust tool for exploring new perceptual spaces and can help identify a lower bound on the number of perceptual dimensions for a given dataset.

3.2 *Scaling, procedure, and simulations*

THIS SECTION first introduces the fundamental concepts of triplets, ordinal embedding algorithms, and triplet accuracy; afterwards, it describes our procedure for dimension estimation and shows results from simulations for validation.

3.2.1 Background: Triplets and ordinal embedding

Triplets reflect stimulus similarities

PSYCHOPHYSICAL SCALING attempts to create a geometric, distance-based representation of perceived (stimulus) similarity. Similarity can be measured by many different experimental tasks of which the *triplet* (or triad) task is reasonably common (Bonnardel et al., 2016; Devinck & Knoblauch, 2012; Haghiri et al., 2020; Lagunas et al., 2019; Toscani et al., 2020; Wills et al., 2009). In the triplet task observers are presented with three different stimuli, usually simultaneously, of which one is called the *anchor*. The observer chooses one of the other two stimuli perceived as most similar (or dissimilar) to the anchor, resulting in the *(anchor, near, far)*-a triplet of stimulus indices. The triplet task comes in different flavours depending on the instructions, e.g. "Which is the odd one out?" (Hebart et al., 2020); or the opposite question "Which appears most central?" (Kleindessner & von Luxburg, 2017). Sometimes observers are presented with more stimuli and are asked to make multiple decisions, as, e.g., in Roads and Love (2021) and Roads and Mozer (2019). From an ordinal embedding perspective, these differences are irrelevant—the responses can be mapped to triplets (interested readers can find details about the mappings in the Supplementary Material A.1).

COLLECTING MORE DATA (triplets) will lead to more accurate results corresponding to a more accurate psychophysical scale in the context of scaling or ordinal embedding. Furthermore, the required number of trials in a triplet experiment also depends on the number of stimuli n—which is known—and the dimensionality d of the stimulus space—which is unknown. As a rule of thumb Haghiri et al. (2020) recommend to use at least $2dn \log_2 n$ triplets. This rule is based on a mathematical proof about the number of triplets required to reconstruct the scale up to a small error (Jain et al., 2016). Because of the proof, we know that often a fraction of the possible $3\binom{n}{3}$ triplets is sufficient to reconstruct the scale and that the number of trials must increase with both the perceived dimension and the number of stimuli.

IN PRACTICE, there are very different triplet-based experiments; whilst their methodological and statistical choices are important, almost always, the choice of stimuli might be even more critical. Some lab-based experiments present less than 100 stimuli in several hundred or few thousand triplets (e.g. Aguilar et al., 2017; Toscani et al., 2020) while crowd-sourced online experiments present up to $50,000$ stimuli in millions of trials (e.g. Hebart et al., 2020; Roads & Love, 2021). Typically, these triplets show distinguishable stimuli, i.e. the differ-

ences are supra-threshold but similar enough for "reasonable" comparisons: variations or changes of material properties or samples from the same domain (e.g., images of landscapes). Otherwise, answering the comparisons might become challenging, or the comparisons measure cognitive associations instead of perception: The question "what is more similar to a tree, the sun or a neuron?" could be judged based on concepts like photosynthesis or the tree-ish look of the neuron's dendrites. Data from such experiments can be embedded into the similarity space but are unlikely to yield insights into the workings of the visual system.

Ordinal embedding methods estimate scales

PSYCHOPHYSICAL SCALES represent the stimuli as coordinates in d dimensions, whose distances should correspond to the perceived stimulus similarity. Ordinal embedding algorithms choose these coordinates $\boldsymbol{\psi}_1, \ldots, \boldsymbol{\psi}_n \in \mathbb{R}^d$ by maximizing the *agreement* between triplets *(anchor, near, far)* and the corresponding distances in terms of Equation 3.1. Similarly, we call triplets where the inequality does not apply *disagreeing triplets*.

$$\text{dist}(\boldsymbol{\psi}_{\text{near}}, \boldsymbol{\psi}_{\text{anchor}}) \leq \text{dist}(\boldsymbol{\psi}_{\text{anchor}}, \boldsymbol{\psi}_{\text{far}}). \tag{3.1}$$

The coordinate estimation requires no stimulus attributes or neighborhoods and, provided enough data, provably recovers metric information up to similarity transformations (translation, rotation, reflection, scaling), and a small error (Jain et al., 2016; Kleindessner & von Luxburg, 2014)—assuming appropriate dimension and distance metrics. The appropriate distance metrics of psychological spaces are actively discussed (for recent discussion, see, e.g. Logvinenko & Maloney, 2006; Love & Roads, 2021), but the standard Euclidean distance is perhaps the most intuitive and the most commonly used; thus we use it here.

ALGORITHMICALLY, ordinal embedding methods optimize coordinates that minimize a *stress*-function on a set of triplets $T = \{(i_1, j_1, k_1), \ldots, (i_m, j_m, k_m)\}$, where i, j, k are stimulus indices whose order correspond to the trial response *anchor*, *near*, *far*. The numerical properties of this stress function (e.g. smoothness) are more desirable for optimization than the agreement count (Equation 3.1). Different ordinal embedding algorithms mainly differ in their stress-function (Agarwal et al., 2007; Jain et al., 2016; Terada & von Luxburg, 2014; van der Maaten & Weinberger, 2012). The algorithm that we use in this work is called *soft ordinal embedding* (SOE Terada & von Luxburg, 2014) and has been shown to result in very accurate reconstruction in a large scale benchmarking

study (Vankadara et al., 2021). SOE's stress function (Equation 3.2) is "soft" in the sense that disagreeing triplets are included based on the size of their squared error. Trivial solutions with all-zero coordinates are prevented by enforcing a minimal distance difference ("$\dots + 1$"). Once a coordinate triplet agrees by this minimal distance it does not increase the stress ("$\max[0, \dots]$").

$$\sum_{(i,j,k)\in T} \max\left[0, \operatorname{dist}(\boldsymbol{\psi}_j, \boldsymbol{\psi}_i) - \operatorname{dist}(\boldsymbol{\psi}_i, \boldsymbol{\psi}_k) + 1\right]^2. \tag{3.2}$$

Equation 3.2 is minimised for coordinates $\boldsymbol{\psi}$ by the Broyden-Fletcher-Goldfarb-Shanno algorithm (BFGS, see Fletcher, 1987). Unfortunately, the optimization is non-convex and sometimes converges to sub-optimal solutions. Thus, each optimization is restarted from ten random initializations, returning the scale with minimal stress.

Triplet accuracy measures the scale's fit

While the stress is useful to maximize the scale's fit on training triplets, it is not a good indicator of the scale's fit. The stress cannot predict how well the scale matches the observer's responses *in general*. A more direct measure of fit is the proportion of triplets that agree in terms of Equation 3.1 with the scale $\boldsymbol{\psi}$, called the *triplet accuracy* (Equation 3.3).

$$\operatorname{acc}(\boldsymbol{\psi}, T) = \frac{1}{m} \sum_{(i,j,k)\in T} \mathbb{1}_{\left[\operatorname{dist}(\boldsymbol{\psi}_j, \boldsymbol{\psi}_i) \leq \operatorname{dist}(\boldsymbol{\psi}_i, \boldsymbol{\psi}_k)\right]}. \tag{3.3}$$

The indicator function $\mathbb{1}_{[\dots]}$ returns 1 for agreement and 0 otherwise, such that $\operatorname{acc}(\boldsymbol{\psi}, T)$ ranges between 0 (full disagreement) and 1 (full agreement).

The triplet accuracy can be calculated either on the same triplets used to fit the scaling algorithm or on a separate set from the same population to test the scale. Thus we can distinguish between *training triplet accuracy* (train accuracy) and *test triplet accuracy* (test accuracy). In contrast to the training accuracy, the test accuracy helps distinguish reasonable from *noisy* responses. Noise in the sensory system, lapses, or other human imperfections, summarised as *judgment noise*, might cause erroneous triplets that disagree with the majority of responses. However, the test triplets likely contain different erroneous triplets; this results in a lower test accuracy that is a better estimate of the general or "true" fit than the (spuriously high) training accuracy.

Cross-validating test accuracy

TRIPLET COLLECTION in a perceptual context is time-consuming. Thus instead of collecting entirely disjunct training and test data, we advocate using a resampling algorithm for a data-efficient approximation of the test accuracy. This resampling-algorithm, *k-fold cross-validation* (compare Hastie et al., 2009), splits the dataset into k equal-sized parts and estimates k scales to calculate k accuracies. For every iteration, $k-1$ different parts are used for training, while the k-th part is used for testing the accuracy. In order to sample more than k accuracies, *r-repeated cross-validation* repeats the k folds on r shuffled versions of the dataset. The mean of the resulting $k \cdot r$ accuracies approximates the test accuracy. Please note that these cross-validated scales are used for accuracy estimation—the "final" scale for visualizing the observer's perception is estimated from all triplets.

Trading off dimensions against accuracy

PSYCHOPHYSICAL SCALES can be seen as parametric models whose parameters are the stimulus coordinates in the perceptual space. Like any parametric model, scales, too, are affected by under- and overfitting if coordinates have too few or too many dimensions. We illustrate this by revisiting the example from the introduction, Figure 3.1. One can imagine triplets encoding the distance relations in the two-dimensional scale. The one-dimensional scale lacks sufficient freedom to capture all triplets; the scale is *underfitting*. For example, red is perceived as more similar to yellow than blue; in the one-dimensional fit, red is farther from yellow than blue. The three-dimensional scale might perfectly represent all triplets but also all the erroneous ones (zero training accuracy)—the scale is *overfitting*. The erroneous triplets, caused by judgement noise, disagree with most triplets and thus with the most accurate scale.

3.2.2 *Our procedure: Testing for accuracy gains*

FOLLOWING the previous considerations of under- and overfitting for different scaling dimensions, a suitable procedure should choose the dimension in which the test accuracy is maximal. However, given the noise inherent in psychophysical data—and thus in scale estimates and accuracies—a purely visual inspection will not do. We require a statistical test for dimensionality based on the triplet accuracy.

Related dimension estimation methods

THE STATISTICS and machine learning literature proposes several dimension estimation methods, but they are unsuitable for analyzing

psychophysical data.

The vast majority of methods in machine learning use metric data (see Camastra & Staiano, 2016, for an overview) , i.e. every data point is described by a collection of numerical features. However, data from perceptual scaling experiments like rankings or triplets only provide the order of stimulus similarities.

ONLY A FEW dimension estimation methods use non-metric data. However, they all require more data than we typically can collect in psychological experiments: The method of Kleindessner and von Luxburg (2015) estimates the dimensionality from information about the k nearest neighbours of each datapoint (e.g. the k most similar stimuli). However, it is not straightforward to calculate the k nearest neighbours from triplets. Additionally, the method's performance highly depends on the number of objects, which are the stimuli in our setting; the authors tested their method with $5 \cdot 10^4$ to $5 \cdot 10^7$ objects, which is far from a feasible stimulus size.

The other non-metric dimension estimation method that we are aware of follows another approach but is similarly difficult to apply in a perceptual setting: Tabaghi et al. (2021) derive a so-called *ordinal capacity* metric that differs between metric spaces, e.g. one-dimensional Euclidean, two-dimensional Euclidean, and four-dimensional hyperbolic space. Calculating this ordinal capacity requires sorting the distances between n stimuli which requires in the limit up to n times more triplets than estimating a (low-dimensional) scale ($\mathcal{O}\left(n^2 \log n\right)$ instead of $\mathcal{O}\left(dn \log n\right)$ with $d \ll n$ Haghiri et al., 2020). In an exemplary case with a 2D scale of 40 stimuli, one requires about 20 times more triplets to determine the ordinal capacity than to estimate the scale. This additional experimental effort—just for determining the dimensionality—is unacceptable.

Test for a significant gain in accuracy

THE ELEMENTS of our procedure are statistical tests between scales with increasing dimensionality d and $d+1$, testing if adding a dimension improves the mean test accuracy μ, with the null hypothesis $H_0^d: \ \mu_{d+1} \leq \mu_d$ and alternative $H_1^d: \ \mu_{d+1} > \mu_d$.

In our procedure, accuracy samples $\mathbf{acc}_d, \mathbf{acc}_{d+1} \in \mathbb{R}^{rk}$ are collected using repeated cross-validation with r repetitions and k folds. By default, we use $r = k = 10$ based on empirical results in the model comparison literature (Bouckaert & Frank, 2004), leading to 100 samples per dimension. The accuracy gain $\mathbf{acc}_{d+1} - \mathbf{acc}_d$, paired by repetition and fold, is evaluated with a two-sample t-test whose test statistic is modified for the use of cross validation.

A STANDARD T-TEST assumes normally distributed and independent samples. While accuracy samples are binomial and thus approximately normally distributed (see Dietterich, 1998, for the argument and see the Supplementary Material A.2 for simulations), their independence is violated by the data overlap in cross-validation folds. For t-tests with k-fold cross-validated datasets, Nadeau and Bengio (2003) proposed the correction factor $\frac{1}{k-1}$ in the test statistic:

$$t = \frac{\text{mean}\left(\mathbf{acc}_{d+1} - \mathbf{acc}_{d}\right)}{\sqrt{\frac{1}{rk} + \frac{1}{k-1}} \cdot \text{sd}\left(\mathbf{acc}_{d+1} - \mathbf{acc}_{d}\right)}. \tag{3.4}$$

We use the test statistic to calculate the probability of obtaining the observed accuracy gain under the null hypothesis that there is no gain, the p-value. We accept the accuracy gain only if the p-value is lower than an acceptance threshold α. Besides the modified test statistic, the calculation is identical to a standard one-sided t-test: The p-value is the probability density of a Student's t distribution with $rk - 1$ degrees of freedom at our t-value, calculated with Equation 3.4.

Sequential testing

THE GOAL is to detect the lowest dimension without an accuracy gain in a predefined range—care has to be taken to apply appropriate multiple-testing corrections to prevent the increased risk of false positives. The tested dimension range is the parameter of interest and depends on the perceptual question, i.e. the experimental task and stimulus.

In a range of dimensionalities, the procedure tests the neighbouring dimensionalities for the alternative hypothesis "gain in accuracy", then returns the lowest gain-providing dimensionality. If no rejection occurs, the intrinsic dimensionality is assumed beyond the tested range. The more neighbouring dimensions are tested, the more likely an erroneous significance occurs (multiple testing problem). The acceptance threshold α should be corrected to compensate for the total number of tests, the best-known correction being the *Bonferroni* method (Bonferroni, 1936). However, despite its charm of simplicity, the Bonferroni correction is known to over-correct α once the number of individual tests increases, in other words, the method has low statistical power (Holm, 1979). In practice, we assume that the corrected α of more than three tests would be too small to detect a "gain in accuracy".

Thus, our procedure uses an improved version of the Bonferroni method with larger statistical power, called the *Holm-Bonferroni method* (Holm's step-down procedure Holm, 1979), to correct the significance threshold α of the neighbouring dimension tests. The Holm-Bonferroni method is more powerful than basic Bonferroni (fewer false-negative test results) and hardly more complicated, but otherwise shares Bon-

ferroni's benefits, such as a lack of distributional and dependence assumptions. While the Bonferroni correction divides the significance threshold α by the number of tests m ($\alpha_{\text{corrected}} = \frac{\alpha}{m}$) the Holm-Bonferroni correction considers the ascending ranking r of all test's p-values: $\alpha_r = \frac{\alpha}{m-r+1}$. The strictest threshold $\frac{\alpha}{m}$ is just used for the smallest p-value, such that fewer "gain in accuracy" tests are erroneously rejected with the Holm-Bonferroni method, although the increase in accuracy usually decreases with increasing dimension.

We became aware that in pharmaceutical studies, there exists a statistically similar problem, the so-called *dose finding problem*: The effect of medicine typically increases with the dose until a certain point where the effect stagnates or decreases—just as the test accuracy in our dimension finding problem. The optimal dose is approached by statistical testing procedures similar to ours (Bauer & Budde, 1994; Budde & Bauer, 1989).

Put together: The dimension-testing procedure

Our procedure detects the dimensionality of triplet data by estimating psychophysical scales and looking for their accuracy peak with a sequential testing scheme:

a) Estimate scales for $d = 1$ to $m + 1$:

1. Estimate and cross-validate k psychophysical scales in d dimensions with SOE, repeat r times on shuffled triplets.
2. Collect triplet accuracies $\mathbf{acc}_d \in \mathbb{R}^{rk}$ from the r-repeated k-fold cross-validation.

b) Test scales pairwise for $d = 1$ to m:

3. Calculate the p_d-value of an accuracy gain $H_d : \mathbf{acc}_{d+1} - \mathbf{acc}_d > 0$ with the Student's t-distribution PDF ($df = kr - 1$) at the t-value of Equation 3.4.

c) Combine the tests for $d = 1$ to m:

4. Accept H_d if $p_d < \frac{\alpha}{m-R(p_d)+1}$,
   R is the rank of p_d.
5. If H_d rejected, return "d dimensions".

d) If no H_d has been rejected,
return "at least $m + 1$ dimensions".

3.2.3 *Simulations: Validating our procedure*

NEW METHODS should always be validated against ground truth, but ground-truth dimensionalities do not exist in "real" psychophysical data. Hence we simulated data from "synthetic" observers. By simulating judgments, we have complete control over all aspects of the data and thus can rigorously assess our statistical procedure.

Generated ground-truth scales

IN ORDER to cover a range of experiments with our simulations, we require ground-truth scales where we can freely choose the number of stimuli or dimensions and re-create comparable variants of the scale. Thus, we sampled the scale's coordinates from normal distributions, which provides us with an infinite amount of ground-truth scales, called *normal scales* in the following. In addition to these normal scales, results of two ground-truth scales inspired by actual psychophysical scales, namely, a circle-like hue (Ekman, 1954) and a helix-like pitch scale (Shepard, 1965) are available in the Supplementary Material A.5.

The normal coordinate distribution $\boldsymbol{\psi}_1, \ldots, \boldsymbol{\psi}_n \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$ has zero mean and an identity covariance matrix (maximum density at the origin). Reproducibility is assured by seeding the pseudo-random generator used to sample from the distribution. Different numbers of stimuli n and dimensions d were chosen to simulate different psychophysical stimuli: Small experiments of $n = 20$ and $d = \{1, 2, 3\}$, medium experiments of $n = 60$ and $d = \{1, 2, 3, 8\}$, and large experiments of $n = 100$ and $d = \{3, 8\}$.

For all these ground-truth scales and simulated triplets, our procedure searched the true dimensionality from 1 to $(d + 2)$, as shown in the following sections.

Simulated triplet judgments

AS IN a lab experiment, we created random triplets of stimulus indices. However, instead of asking observers to judge, we calculated judgments from distances in a ground-truth scale plus judgment noise. For each ground-truth scale, we created multiple datasets with a different number of trials.

Comparing validation results from different ground-truth scales requires "a common currency" of the evaluated metrics. However, the quality of scale estimates and, thus, our dimension estimates depend not on the absolute number of trials but on the ground-truth dimension d and stimulus number n. Therefore comparable trial numbers were calculated with the scaling factor λ and the $\lambda dn \log n$- formula (Haghiri et al., 2020), based on mathematical proofs in the computer

science literature. We used the natural logarithm (an arbitrary decision) and varied λ to define three different dataset sizes: The "minimal" dataset ($\lambda = 2$), the "moderate" dataset ($\lambda = 4$), and the "generous" dataset ($\lambda = 8$). In addition, a sample of $10,000$ triplets accompanied every triplet dataset to approximate the *noise ceiling*, the best possible generalization accuracy considering the fraction of triplets that became incompatible through the (simulated) judgment noise.

We created the dataset's triplets by random sampling of three distinct stimulus indices, i, j, k, that were judged from the Euclidean distances between ground-truth positions $\boldsymbol{\psi}_i, \boldsymbol{\psi}_j, \boldsymbol{\psi}_k$ as

$$\begin{aligned} &(i,j,k), \quad \text{if } \operatorname{dist}(\boldsymbol{\psi}_j, \boldsymbol{\psi}_i) + \epsilon \leq \operatorname{dist}(\boldsymbol{\psi}_i, \boldsymbol{\psi}_k) \\ &(i,k,j), \quad \text{otherwise.} \end{aligned}$$

At every judgment, the noise component ϵ was sampled from a normal distribution $\mathcal{N}(0, \sigma^2)$ to simulate judgment noise as in similar simulation studies (Aguilar et al., 2017; Devinck & Knoblauch, 2012; Haghiri et al., 2020). The normal noise models observers that misjudge closely perceived similarities more frequently, i.e. visual similarity judgments between three different red apples should be less consistent than two red apples and one green pear. Three noise levels σ were defined as low ($\sigma_{\text{low}} = 0.5$, e.g. controlled lab experiments), medium ($\sigma_{\text{med}} = 1.0$), or high judgment noise ($\sigma_{\text{high}} = 2.0$, e.g. online experiment); the interested reader can find a visualization of these levels in the Supplementary Material A.3. We rescaled these noise levels according to distance's spread to maintain comparable signal-to-noise ratios across different simulation settings and approximately match the triplet accuracy range in corresponding human datasets.

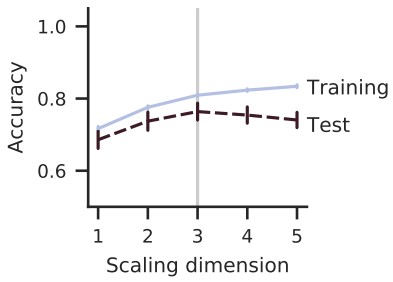


Figure 3.2: Comparison of training and test triplet accuracies for different embedding dimensionalities (#triplets $= 2947, \lambda = 4,$ noise $=$ med). The triplets are simulated with medium judgment noise from an artificial 3D scale with 60 normally distributed points. The training accuracy increases with the dimensionality, but test accuracy peaks at the ground-truth dimensionality (vertical line). The standard deviation between cross-validation folds (error bars) is higher for the test accuracy.

Accuracy peaks at the ground-truth dimension

The first results we look at are accuracy-by-dimension graphs as the underlying metric of our procedure, whose key idea is to identify a test-accuracy peak at the ground-truth dimensionality. The following representative results use datasets with the 3D-normal scale ($n = 60$). Datasets with varied dimensionality and number of stimuli but comparable results are shown in the Supplementary Material A.6.1.

The accuracy on training triplets in Figure 3.2 increases with the embedding dimensionality as the scale fits more and more triplets. However, the accuracy on test triplets peaks at 3D, the ground-truth dimensionality indicated by the vertical line, and shows that the scale is overfitting to noisy triplets for higher dimensionalities.

The effect of dataset size on the test accuracy is negligible for scales of ground-truth dimensionality but not for scales with higher dimen-

sionalities: Figure 3.3 (right) shows a pronounced accuracy peak for the small dataset, but the accuracy for larger datasets converges to the noise ceiling (horizontal line). The dataset size also influences the slope of the training accuracy, such that the training accuracy of a large dataset just minorly increases beyond the ground-truth dimensionality. This reduced slope of training and test accuracies reduces their difference by increasing the dataset size.

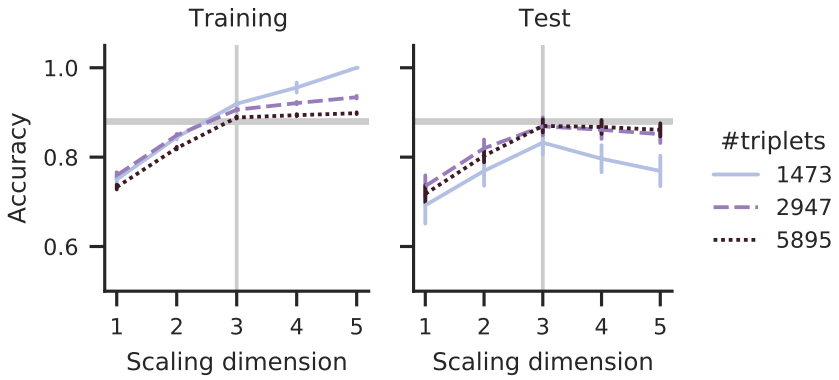


Figure 3.3: Accuracies for different dataset sizes and low simulated noise. Dataset sizes affect the test accuracy peak only mildly; it has a stronger influence on how much accuracy increases (train; left panel) or decreases (test; right panel) after the ground-truth dimensionality (vertical line).

In contrast to dataset size, noise severely reduces the noise ceiling and thus the achievable accuracies (Figure 3.4). Additionally, noise flattens the accuracy graph, which thus shows less pronounced peaks, which might reduce the precision of dimension estimates.

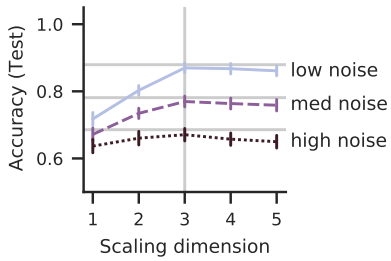


Figure 3.4: Test accuracies for simulated noise with different signal-to-noise ratios (#triplets $= 5895, \lambda = 8$). The noise reduces the best-possible accuracy (noise ceiling, horizontal lines) leads to flat accuracy graphs. The high noise accuracy shows no peak at the ground-truth dimensionality (vertical line).

Estimated dimensionality is conservative

THE FOLLOWING RESULTS are our procedure's dimensionality predictions, based on statistical tests for a gain in accuracy.

The statistical test's p-values below $\alpha = .05$ indicates a significant gain in accuracy by adding another dimension to the scale. Figure 3.5 shows these p-values along with the predicted dimensionality at the first rejection of the gain hypothesis (red line) for multiple noise levels and dataset sizes. The accuracy peaks were reliably detected at the ground-truth dimensionality even for settings where the peak is barely visible (compare high noise graph in Figure 3.4). Only for one small dataset was dimensionality underestimated (left panel middle row in Figure 3.5).

Across all 81 simulations of normally distributed ground-truth scales, our procedure estimates the correct dimensionality 73% of the time. All incorrect predictions underestimated the ground-truth dimensionality. These underestimates occurred more frequently for small datasets, high noise, or large ground-truth dimensionality. The individual dimensionality predictions are summarized in the Supplementary Materials A.6. Please note that 73% correct might appear low; however, this

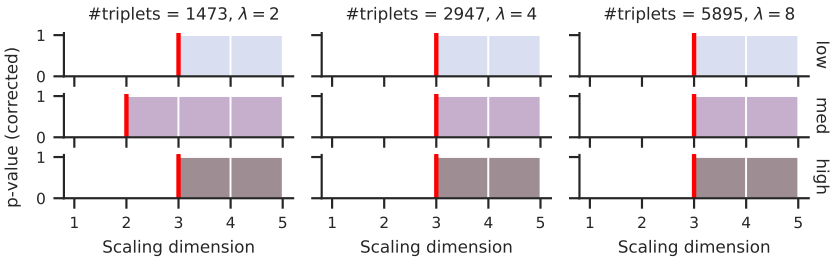


Figure 3.5: p-values of statistical tests to detect accuracy gains by adding a dimension to the estimated scale from simulated triplets of a 3D ground-truth scale with 60 stimuli. Colors and vertical order match the noise levels of Figure 3.4. The predicted scale dimensionality (red lines) matches the 3D ground-truth in most settings; for more than 3D, the accuracy gain was always rejected ($p > .05$).

is only a reflection of the fact that we used very challenging simulation conditions with (sometimes) just the minimal amount of data and substantial noise. Our results clearly show the considerable influence of noise on correct dimensionality estimation. If we only consider low-noise settings, 93% (25 of 27) dimensionalities were predicted correctly, even including the small datasets. The incorrect predictions were with datasets of few stimuli given the ground-truth dimensionality (3D and $n = 20$; 8D and $n = 60$). This result indicates that the stimulus number might be another factor affecting the robustness of dimensionality estimation; this factor is common to all dimension estimators and is addressed in the final discussion.

Repeated simulations show reproducability

In the previous sections, we showed single runs of our method on various datasets to investigate the relevant parameters. Here, we repeat the procedure 100 times on the same dataset to evaluate the robustness of the procedure. This section shows results for triplets from the 8D normal scale ($n = 100$), while comparable results on different datasets are available in the Supplementary Material A.6.1.

The procedure consists of statistical tests to detect if adding a dimension increases the accuracy (compare p-values in Figure 3.5). We expect significant increases until the dimensionality equals the ground-truth scale's dimensionality. Figure 3.6 depicts how many of our procedure repetitions violate this expectation. The left plot shows that almost no incorrect accuracy gain was detected, which is expected from the conservative multiple-testing correction and the robust decrease of test accuracy after the ground-truth dimensionality in the previous plots. The dark colors in the right plot indicate that the statistical test rejected the accuracy gain hypothesis for some scaling dimensions lower than the ground truth; the procedure underestimated the dimensionality. These underestimates occur more frequently for high noise settings (Figure 3.6, right) and for small datasets and high ground-truth dimensionality (see Supplementary Materials A.6). Overall, the repeated runs of our method confirm the large influence of the noise

magnitude on scaling accuracy and dimensionality under-estimation.

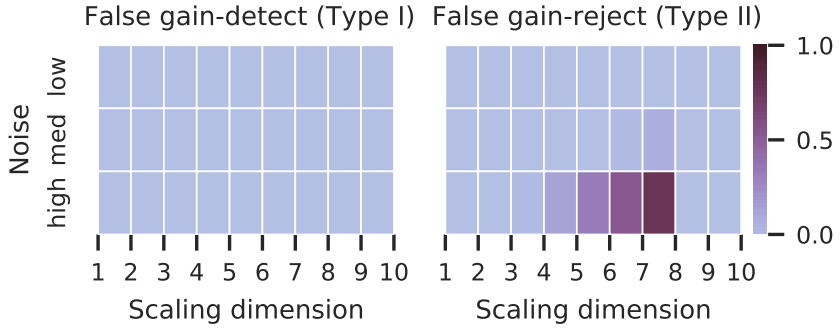


Figure 3.6: Unexpected rejections and detections of our neighboring-dimension tests for repeated simulations of 8D normal scales ($n = 100$) with medium triplet size ($\lambda = 4$). With noisy triplets, the accuracy gain is rejected even before the ground-truth dimension leading to lower-bound estimates of the dimension.

Summary

OUR PROCEDURE reliably identifies the ground-truth dimensionality in the simulated datasets if enough trials were collected and the noise is low. Collecting more trials can only partially offset the noise; thus, the focus should be on controlling judgment noise through control measures in the experiment. However, our procedure identifies a lower-bound dimension estimate even in the worst-case conditions of high noise and few trials. This dimensionality-underestimation of sparse and noisy data is—in our opinion—preferable behavior because it provides the user with a more straightforward explanation. Such worst-case conditions are easily identified by monitoring the train and test accuracies. A large gap between training and test accuracy and the considerable variation of accuracies within cross-validation folds indicates that more trials should be collected, while low accuracy indicates considerable noise.

3.3 *Dimensionality of human data*

RESULTS of the previous section showed that our procedure could predict the dimensionality of *simulated trials*. However, the intended application of our procedure is dimension estimation on *behavioral trials* from psychophysical experiments. Thus we also investigated behavioral datasets. In contrast to simulations, behavioral data has no ground truth but just more or less evidence about the "correct" dimensionality.

The section starts with hue-triplets as a sanity check, where we expect and find a two-dimensional representation and continues with two other datasets where the true dimensionality is less evident. Its prediction is—perhaps—somewhat surprising.

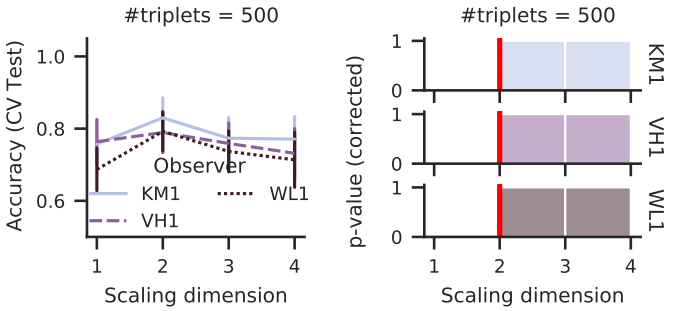


Figure 3.7: The clear estimation of two perceived hue dimensions matches the color wheel representation even if the original data were the four-dimensional ratings of Bosten and Boehm (2014).

Hue: Verified expectations

Color is a natural testbed of multi-dimensional perceptual spaces. One property of colors, the hue, might be represented with a color wheel that requires two euclidean dimensions even though the corresponding physical parameter is one-dimensional (wavelength of light). The hue similarity is typically collected in rating experiments; however, we computed triplet trials from the ratings.

The ratings of 36 different hues were collected by (Bosten & Boehm, 2014) from 18 observers. Every hue was presented in three trials such that every observer answered 108 trials. On each trial, a test patch was presented, and the observer rated (from 0 to 9) the similarity of the patch's hue to red, yellow, green, and blue; thus, the observer supplied four numbers on each trial. For example, observers experience a violet hue with red and blue but with little yellow and green. One might think of these ratings as samples in a 4D space with a red, yellow, green and blue axis that can be used to judge hue triplets. Per triplet, we randomly selected a target and two other hues and calculated the Euclidean distance of the corresponding 4D ratings to judge which hues were more similar. This way, the triplets involve—to a certain degree—the behavioral noise in contrast to the simulated noise in the previous section.

Figure 3.7 shows our procedure's estimates for hue triplets of three arbitrarily picked observers. For all of them, our procedure suggested a two-dimensional scale that fits very well with the assumed color-wheel representation of the hue. We note that neither data collection nor triplet sampling involved a two-dimensional bias; instead, the data were collected as 4D ratings.

Slant-from-texture: Revealed influences

Another common, but less apparent, percept of interest is the slant of angled textured planes (Rosas et al., 2004, 2005, 2007). The common assumption is that slant and angle are single-dimensional and relate monotonically.

Here we used a triplet-dataset of slant-stimuli by Aguilar et al. (2017), where observers compared three dot textured planes ("polka dots" by Rosas et al., 2004) per trial, which varied in 8 angles. In total 840 triplets were collected, such that each triplet shows angles (left < anchor < right). The ordering is a restriction of the MLDS algorithm (Knoblauch & Maloney, 2012b) that was used to estimate the scales in the original publication (Aguilar et al., 2017). Using MLDS, they could only consider 1-dimensional, monotonic scales. Following up, Haghiri et al. (2020) re-analyzed these triplet data with ordinal embedding algorithms to relax the monotonicity assumption and observed a surprising "dip" of slant.

Here, we even further question the assumption that the resulting scale has to be 1D by applying our procedure. Perhaps surprisingly, our procedure predicts multi-dimensional scales for some observers: The one-dimensional scale was suggested for three out of eight observers (Figure 3.8); the other scales were estimated as two (four observers) or even three-dimensional (one observer).

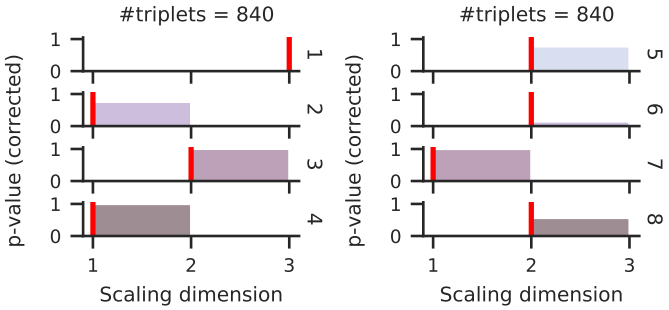


Figure 3.8: The optimal scaling dimension of slant varies between 8 observers (rows), according to our procedure—just three observer's scales were one-dimensional as expected. p-values below $\alpha = .05$ indicate rejection of the H_0 ="No accuracy gain by adding a dimension".

A 2D or 3D slant scale is contra-intuitive given that only the angle varied in the experiment. As there is no ground-truth dimensionality, this surprising result may question our procedure's reliability or indicate that additional dimensions were observed. The reliability of our procedure was thoroughly tested in simulation experiments, and in no condition—not once—did our procedure overestimate the dimensionality; even in deliberately poor datasets, the dimensionality was underestimated, so there is no reason to believe that our procedure failed for this slant-dataset.

The alternative explanation, additional perceived dimensions, could be related to the stimulus design. Even though the independent slant variable is one-dimensional, the stimulus is a high-dimensional image of a dot pattern. Observers had no direct access to the slant angle. However, they must have inferred it from one or several stimulus properties, e.g. changes in the size, width, aspect ratio or density of the texture elements. Perhaps some of the observers switched be-

Figure 3.9: Three distortions of the same landscape image, created with the Eidolon Factory and used as stimuli in the lab experiment of Haghiri et al. (2019).

tween the cues they used or changed their cue-combination strategy as a function of the angle. Clearly, without further experiments, this issue cannot be settled. However, to us, it indicates that one should always consider multi-dimensional scales if only to confirm that a presumed 1D relationship is indeed 1D.

Eidolon's distortions: Correlated parameters

THE THIRD DATASET uses high dimensional stimuli, as shown in Figure 3.9, distorted versions of landscape photography that differ in most pixels. However, the distortions are defined by three parameters *reach, grain, and coherence* of the Eidolon Factory (Koenderink et al., 2017). Triplets of 100 such distorted stimuli of the same landscape photography were generated by Haghiri et al. (2019) for their laboratory experiment. Their observers were asked 6,000 random triplet questions and responded to almost all of them (1st observer, 6,000 responses; 2nd, 5,996; 3rd, 5,999).

From the three parameters of the Eidolon Factory, one might expect three perceived dimensions, but previously Haghiri et al. (2020) observed a peak in mean-accuracy for two-dimensional scales. Our procedure also predicts a 2D scale for two observers and a 3D scale for one observer (Figure 3.10). Again, from our simulations, we believe we are unlikely to overestimate the perceptual dimensionality. Furthermore, we observe relatively high accuracies and only a small gap between train- and test accuracy, indicating that the noise in the dataset is relatively low (and thus, our dimension estimates are very likely correct).

Observing two perceptual dimensions given the three perturbation parameters of Eidolon means that multiple image generation parameters lead to similar percepts, i.e. at least two observers did not perceive the (subtle) differences between all the perturbations.

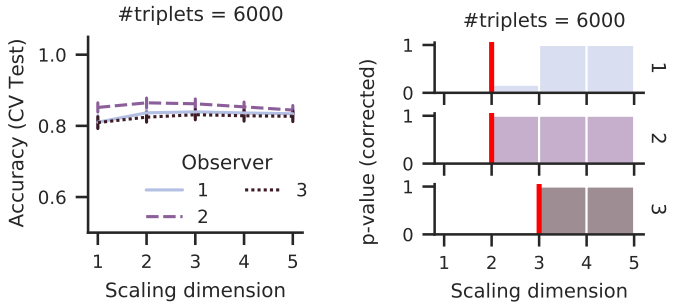


Figure 3.10: The optimal scale for the eidolon triplets is two-dimensional, which supports the observations of Haghiri et al. (2020). This result counters the first intuition of a three-dimensional scale because the stimuli were created with a three-dimensional distortion algorithm.

3.4 *Discussion*

We propose a procedure to estimate the appropriate dimensionality in psychophysical scaling. Our procedure is based on model selection and statistical hypothesis testing to provide a more objective decision than previous approaches. We show in simulation studies that this procedure can recover the ground-truth dimension and produces conservative estimates in noisy settings where "classical" dimension estimators typically overestimate the dimensionalities.

Using three existing behavioral datasets, we showed the use of our procedure in practice; in the case of color, we confirmed the expected 2D embedding: the hue or color circle. For the slant-from-texture and eidolon experiments, however, our procedure uncovered higher (slant-from-texture) or lower (eidolon) embedding dimensions than one might have predicted based on the number of explicitly manipulated variables in the experiments (one and three, respectively).

3.4.1 *Robust perceptual dimensionality estimation*

The robustness of our procedure's predictions was validated in multiple simulation experiments. These validations are essential because one can not compare with ground truth, and errors in the procedure would be taken for reality. However, the validity of simulation-based validations is based on assumptions that link the simulations with the behavioral studies to which our procedure should be applied. In psychophysical scaling, we expect few observers to judge many trials in a well-controlled lab environment. Observers are analyzed separately, so responses are consistent, low-dimensional, and of low perceptual noise (Gaussian distributed).

If an experiment is of this type, we have shown the reliability of our procedure and are confident that it returns accurate scale and dimension estimates.

3.4.2 *Lower-bound estimates from ill-defined data*

TYPICAL PSYCHOPHYSICAL DATASETS are known for their high level of control. Yet, circumstances—e.g. large-scale online experiments with little to no control over the screen, room, attention, noise levels etc.—can lead to too few trials or too much noise, i.e. too many random responses. These data deficits reduce the scale's accuracy. Ultimately low-quality data lack the information required to reconstruct the original scale. Our simulations showed that large noise in the data is the most detrimental factor: Doubling the noise cannot be compensated for by doubling the amount of data. In low data quality scenarios—large noise—our simulations show our procedure to err on the conservative side, i.e. to propose a lower-dimensional scale than ground truth. Again we believe this to be a feature rather than a bug in the context of inferring perceptual dimensions.

Recognizing whether a dimension estimate is lower than expected because the perceptual space is low dimensional or because the dataset is too small and noisy is obviously essential. Two metrics need to be inspected to decide between these two possibilities: First, the maximum test accuracy and second, the difference between training and test accuracy. Low-noise settings show a maximum accuracy of about .9; accuracies below .7 are critical and indicate high noise, thus an increased risk of dimensionality underestimates. Estimates derived from small datasets show low accuracies, too, but are easier to detect by comparing the number of triplets with, e.g. the $2dn \log n$-rule (Haghiri et al., 2020), for different hypothesized dimensionalities d. Additionally, a large difference between train and test accuracy ($\gg 0.1$) can also indicate a lack of data.

The lack of data can be resolved by running additional lab sessions, but reducing the noise might be more difficult. Typical strategies to reduce the noise involve a well-controlled lab environment (Haghiri et al., 2019; Wichmann & Jäkel, 2018), varying the task and extending training sessions (e.g. triplets instead of Likert ratings; Demiralp et al., 2014) or post-hoc data cleaning (e.g. dropping blocks where repeated trials disagree; Lagunas et al., 2019).

3.4.3 *Estimates of high-dimensional spaces*

IN RECENT YEARS there is a trend to investigate perceptual space in large scale online experiments using stimuli like object photographs (Hebart et al., 2020; Roads & Love, 2021). In these studies, the data of very many observers are pooled and then jointly embedded.

The perceptual spaces identified in the above studies tend to be rather high-dimensional. However, this high dimensionality might not necessarily reflect the "the internal human object perception space"

but might instead be (partially) an overlapping *super-space* (Carroll & Chang, 1970), composed of the multiple observer's (cognitive) decision criteria.

This possibility of obtaining "compositional super-sets" from such experiments makes it difficult to reconstruct individual perceptual spaces from representational accuracy. It is thus not an intended application of our procedure.

FURTHERMORE, we would like to highlight the general difficulty of estimating dimensions if their value is high. Intuitively, a space is d-dimensional if its points "cover" a (small) cube of d dimensions. However, the number of points that is needed to "cover" a d-dimensional cube grows exponentially with the dimension d. To see this, imagine 10 data points that "cover" the 1-dimensional interval $[0, 1]$, for example the grid points $0.1, \ldots, 0.9, 1$. To cover a 2-dimensional cube similarly well, we would already need $10 \times 10 = 100$ data points. In general, to "cover" a D-dimensional cube we would need on the order 10^d many points. This fact makes it very difficult to estimate the dimension from a sample of points when D is large. It is pretty much impossible to have enough sample points to be able to distinguish between, say, a space of 50 versus a space of 51 dimensions: our sample points will neither "cover" a cube of 50 nor of 51 dimensions, making each such estimate (or corresponding test) utterly unreliable. A more formal argument for the difficulty of estimating high dimensions can be found in Block et al. (2021). Consequently, while it is well possible in psychophysics to discriminate a 2D from a 3D space, it seems pretty much impossible to discriminate between, say, 50D vs 51D or 50D vs 60D. Even in a setting with very low noise, the high-dimensional scenario would require a prohibitively large number of data points (stimuli) and triplet trials for dimensionality estimation. In psychophysics, it might often be better to avoid high-dimensional spaces from the outset by using well-designed stimuli that observers judge by a few criteria.

4 Quality and stability of ordinal embeddings

I beseech you, in the bowels of Christ, think it possible you may be mistaken.
— Oliver Cromwell

It is a fundamental principle of modeling that a fitted model can only be interpreted if the quality and variability of the parameters are validated. Ordinal embeddings as models of perceived similarity are no different. This chapter explores how ordinal embedding methods behave under incomplete and noisy data and how the quality and stability of the scale can be determined and their variability quantified. The content of the chapter corresponds to an unpublished manuscript. It is mainly my work with contributions to ideation, data analysis, and writing by Felix Wichmann and Ulrike von Luxburg.

Abstract

Human perception of colors, materials, or even objects is often modeled as similarity in a perceptual space. Multidimensional scaling methods represent this perceived similarity of stimuli as distances between points. An intuitive task for assessing this similarity in a scientific experiment is the triplet, a comparison of the form "Is stimulus j or k more similar to i?". Ordinal embedding algorithms can infer the scale points from a random subset of triplet responses. In practice, however, it is often unclear how precise the inferred scale is and whether any scale differences between conditions or subjects are not simply attributable to the inevitably noisy responses. Here we investigate the quality of ordinal embedding estimates from psychophysical data and present a probabilistic approach to model the scale's stability. Using simulation studies, we show that the validation triplet error is not a perfect predictor of the scale's correctness, i.e. whether it reconstructs a typically unknown ground truth. Noisy responses lead to inconsistencies that cannot be resolved and increase the triplet error, no matter how well the embedding describes ground truth; however, the embedding itself can compensate for these inconsistent responses by collecting additional data. Nonetheless, if the error on the training set is subtracted from the validation error, the result is a metric that is corrected for the inconsistent triplets and, therefore, is a useful quality metric. Additionally, we observe that scale ensembles from bootstrapped triplets tend to show a distorted view of the local stability because similarity comparisons provide only relative distance information. The resampled scales' origin, scaling, and rotation must be aligned, which can severely distort the variability estimates. We present a probabilistic model to estimate the variation at individual scale points. The idea behind the model is to "wiggle" the points so that the triplet error changes, indicating the scale's stability. These stability measures can be used to quantify the scale's variability and visualize it as error bars or regions, as we show in application examples with behavioral datasets. These quantifications and visualizations empower scientists to distinguish behavioral effects from mere modeling artifacts.

4.1 Introduction

For more than a century, psychophysicists have been measuring so-called psychophysical scales to describe the perception of our environment objectively. These scales provide numerical intensities to describe how bright the light appears (Aguilar & Maertens, 2020; Obein et al., 2004; Stevens, 1957, e.g.), a 2D "map" of color or material patches arranged by their perceived similarity (e.g., Bonnardel et al., 2016; Logvinenko & Maloney, 2006; Wills et al., 2009), or even high-dimensional coordinates to describe the mental representation of objects (e.g., Hebart et al., 2020; Roads & Love, 2021). In each case, the distance between points on the scale describes the perceived similarity of stimuli as determined in psychophysical experiments.

Most psychophysicists analyze scales primarily by visualizing and interpreting distances between stimuli, conditions, and observers—they draw conclusions about scientific theories "chi-by-eye" (Press et al., 2007). For example, Bonnardel et al. (2016) observe in their 2D color scale that "[t]he configurations are not exactly circular but oval. This matches the elongation displayed when the NCS stimuli are displayed in CIE-L* a* b* space", Hebart et al. (2020) interpret the high-dimensional object scale where "[t]he global similarity structure seems to highlight the well-known distinctions of animate vs inanimate and natural vs man-made", and Sauer et al. (2024) state about scales of perceived lens distortions that "[t]he remaining differences between subjects' scaling functions (for example, differences in the relative influence of *Add* power) could result from individual differences in perception or behavior".

However, we cannot assume that any scale accurately represents the "true" perceived similarity due to the limitations of data collection and modeling. The various stages of stimulus reception and processing up to the motor response are subject to a certain amount of noise—the observer's responses are rarely deterministic and occasionally are contradictory, emphasizing the necessity for stochastic models. Signal Detection Theory, for example, models the stimulus discriminability using *probabilities* for stimulus and response presence or absense (Green & Swets, 1966; Tanner Jr. & Swets, 1954). In addition to noisiness, experiments are limited in the data that can be collect within a reasonable time frame and analysis have to include assumptions due to technical limitations. For example, for scale fitting and visualization, it is convenient to model stimulus similarity with a Euclidean distance, but this does not necessarily correspond best to the observer's intrinsic similarity measure (cf. Jäkel et al., 2008; Logvinenko & Maloney, 2006; Shepard, 1987; Tversky & Gati, 1982) All these confounds—noise, incomplete data, fitting, and assumptions—might cause the empirical

scale estimate to deviate from the unknown inner space. Before we can determine this potential deviation, we need to look at how perception is measured.

EXPERIMENTAL MEASUREMENT of perceived similarity that we consider in this work comes in the form of ordinal comparisons between stimulus pairs. These comparison-based tasks apparently are particularly intuitive for observers (Aguilar et al., 2017; Demiralp et al., 2014) and are becoming increasingly popular (e.g., Aguilar & Maertens, 2020; Charrier et al., 2007; Fleming et al., 2011; Hebart et al., 2020; Knoblauch et al., 2020; Roads & Love, 2021; Wills et al., 2009). A concrete example is in the triplet or triad task, where observers judge if stimulus j or k is more similar to i (Torgerson, 1952). From multiple judgments, ordinal embedding algorithms estimate the points of the scale so that the order of Euclidean distances corresponds to the observer's triplet judgments. In the limit, ordinal embedding methods can recover metric information except for similarity transformations (rotation, translation, scaling; Arias-Castro, 2017; Jain et al., 2016; Kleindessner & von Luxburg, 2014), whereby the required number of triplets increases with the number of stimuli N and the dimensionality of the psychological space D (the error is bounded by $\mathcal{O}(DN \log N)$, see Haghiri et al., 2020; Jain et al., 2016).

The literature proposed various methods to evaluate ordinal embeddings. The goodness-of-fit can be tested statistically on the training data via the agreement between triplet responses and scale estimate (cf. Knoblauch & Maloney, 2008, 2012a; Maloney & Yang, 2003). Predicting responses to the (unknown) ground truth is a typical measure of scale quality. Some works measure this quality as the likelihood on a set of validation triplets (Roads & Mozer, 2019), others with the triplet accuracy, i.e., the proportion of triplet responses consistent with the scale's distances (Haghiri et al., 2020). The *noise ceiling* bounds this prediction accuracy in practice. No estimator can predict intrinsically inconsistent responses, bounding the maximum accuracy.

All these measures are well suited for comparing different scales based on their fit or predictive power (e.g., Künstle et al., 2022a; Maloney & Knoblauch, 2020; Roads & Love, 2021), but are challenging to interpret in absolute terms. The (validation) triplet accuracy is the most interpretable and linked to observer responses. However, the noise ceiling makes it difficult to interpret the error in absolute values (cf. Künstle et al., 2022a) and, above all, to infer the quality of the estimated representation.

CONTRARY to the fit, the stability is usually determined directly in the scale space. For example, Knoblauch and Maloney (2008) sug-

gest a *bootstrapping* procedure to estimate the variability[1] of the scaling method *MLDS*. Triplet responses are sampled parametrically to obtain a distribution of scale estimates and their standard deviation. The standard deviation can be calculated across scales because MLDS constrains scales to 1-dimensional monotonic functions starting at the origin. In contrast, arbitrary D-dimensional scales must be aligned with similarity transformations. For example, the Procrustes method, which minimizes the mean Euclidean error between the scales (Gower, 1975), can be used for this purpose (cf. Haghiri et al., 2020; Lohaus et al., 2019; Sauer et al., 2024)[2]. Lohaus et al. (2019) showed that this error increases with noise in the triplet responses and decreases with additional triplets, expected from a variability measure.

[1] In the machine learning literature, stability and variability sometimes describe slightly different model properties. However, we use them reciprocally to describe the variation when estimating the same ground truth pattern empirically.

[2] Our comparison-based scaling methods share the alignment problem with their precursor, Multidimensional Scaling (Kruskal, 1964a, 1964b). Many of the approaches discussed here can be found in similar form in MDS literature (cf. Gronau & Lee, 2020; Jacoby & Armstrong II, 2014; Weinberg et al., 1984).

Probabilistic models take a different approach, directly learning the scale and its uncertainty as a distribution. The Hefner model describes each stimulus in the scale by an isotropic normal distribution (Hefner, 1958) so that the distances of the mean values express the perceived similarity relative to the variation. Historically, mathematical estimates of multidimensional mean and variance were feasible only through simplifications and independence assumptions (Ramsay, 1969; Zinnes & MacKay, 1983). However, modern probabilistic modeling methods with variational inference enable even larger scales to be described efficiently using normal distributions (Muttenthaler et al., 2022; Roads & Love, 2021). In general, however, it is rather doubtful that the uncertainty is symmetrical and should, therefore, be described or approximated with a normal distribution. Models inferred with Monte Carlo samplers allow a much more flexible description of the scale's variability (Lohaus et al., 2019; Roads & Mozer, 2019; Westfall & Lee, 2021). Like in bootstrapping, an alignment of samples must be enforced, typically using a carefully selected prior (cf. "The Identifiability Problem" Gronau & Lee, 2020). Practical applications of Monte Carlo sampling can be found in active query methods that approximate independence (i.e., assuming independence of stimuli) to select triplets that maximize information gain (Heim et al., 2015; Tamuz et al., 2011).

Despite this plethora of putative quality and stability metrics, it is still unclear whether a scale estimate of experimentally collected triplets describes perceived similarity. The lack of validation experiments partly causes this unfavorable situation. In addition, researchers in cognitive science are particularly interested in whether local differences in the perceptual space are meaningful, i.e., if differences in scale values are larger than their expected variation due to estimation uncertainty. In this work, provide metrics and methods for determining whether we can scientifically trust the interpretation of a comparison-based psychophysical scale.

Our contribution is as follows:

1. We show in simulation studies that neither goodness-of-fit nor prediction is a sufficient quality metric. Instead, the difference in triplet errors on validation and training data indicates the recovery error of the scale.

2. We point out the limitations of bootstrapping methods, where the alignment of scales distorted the variability locally.

3. We develop a probabilistic model to estimate the variability from an existing scale and its training triplets. The model's intuition is to measure variability by "wiggling" the scale's points only as far as they still match the triplets.

4. On behavioral datasets, we apply this quality metric and variability model to gain scientific insights.

4.2 *Quality metrics of ordinal embeddings*

As scientists, we expect a scaling model to describe the observer's perception correctly. Often, this "correctness" is referred to as the *quality* of the scale and is measured by the predictive accuracy of the scale. In this section, we use simulation experiments to investigate whether we can infer "correctness" from "predictions" under realistic conditions. In addition, we ask how reproducible the estimates are with different data sets of the same ground truth, i.e. whether the scale is stable.

4.2.1 *Experiment simulations and scaling algorithms*

In our simulation study, we start with hypothetical ground-truth similarities in place of the subject's perception that is unknown in real experiments. With this ground truth and some noise, we answer a random set of triplets and estimate a scale with ordinal embedding algorithms. We can then evaluate this scale with quality metrics based on the ground truth.

Simulations

The hypothetical ground-truth perception mimics human color perception. This relatively well-studied area of human perception provides an intuitive testbed for 1D and multidimensional scaling simulations.

We simulate the ground truth of 1D lightness perception and 2D hue perception by first selecting points $\boldsymbol{\psi}_1, \dots \boldsymbol{\psi}_N$ from the *CIELUV*

color space (the interested reader can find an introduction into colorspaces in Fairchild, 2013), and then adding samples from a statistical distribution representing perceptual noise. The lightness scale is built by perceptually equidistant grey points between $\psi_1 = L^*_{\text{min}} = 0$ and $\psi_10 = L^*_{\text{max}} = 100$ with $U^* = V^* = 0$. Noise increases with L^* to simulate Poisson distributed perceptual noise in the perceptual domain. It is simulated by adding samples from a Gaussian distribution $\mathcal{N}(0, \sqrt{L^*})$ to the scale. The hue scale is specified as equidistant points on a circle of radius 70, center $U^* = 30$, $V^* = -35$, and $L^* = 50$. The CIELUV space specifies the coordinates L^*, U^*, and V^* such that their Euclidean distance 1 corresponds to a just-perceptible color difference. Therefore, we add to each point a sample from an isotropic Gaussian distribution $\mathcal{N}((0,0)^T, (10,10)^T)$. We multiply the noise variance with a scaling parameter ω (default: $\omega = 1$) to experimentally vary the standard noise level.

We sample the dataset T randomly from all possible triplets. To keep results comparable between simulations, we parametrize the dataset size with a size parameter λ (default: $\lambda = 2$); the number of triplets is calculated with $\lambda DN \log_2 N$ via the number of points N and ground-truth dimensions D. The response to "is the jth or kth stimulus more similar to the ith" is simulated by comparing the Euclidean distance $d(\cdot,\cdot)$ on perturbed ground-truth $\boldsymbol{\psi}'$, created by adding a sample of the noise distribution to the ground-truth points. The order of indices in the triplet (i,j,k) implies that $d(\boldsymbol{\psi}'_i, \boldsymbol{\psi}'_j) < d(\boldsymbol{\psi}'_i, \boldsymbol{\psi}'_k)$.

Ordinal embedding algorithms

Psychophysical scales can be estimated by ordinal embedding algorithms, searching for the points that most likely agree with the triplet responses (Haghiri et al., 2020).

The most common method in psychophysics, MLDS, uses logistic regression to predict the response to triplets; the weights of the regressor are the scale (Knoblauch & Maloney, 2008, 2012a; Maloney & Yang, 2003). Therefore, the algorithm cannot use arbitrary triplets and constrains the scale to be 1D, monotonically increasing. Much more flexible are ordinal embedding algorithms, which can also estimate multidimensional scales with arbitrary triplets (Agarwal et al., 2007; Jain et al., 2016; Tamuz et al., 2011; Terada & von Luxburg, 2014; van der Maaten & Weinberger, 2012). Soft ordinal embedding (SOE; Terada & von Luxburg, 2014), in particular, showed high training and test accuracy on various datasets (Künstle & von Luxburg, 2024; Vankadara et al., 2021), which is why we use it in this work. In general, however, we intend the developed methods of this work to function largely independently of the choice of embedding algorithms.

SOE searches the points $\hat{\boldsymbol{x}} \in \mathrm{R}^D$ by minimizing the sum of squared errors between the "less similar" and "more similar" distances of each triplet (i, j, k). In our experiments, we use the Python implementation in the *cblearn* toolbox (Künstle & von Luxburg, 2024). We chose the dimensionality D of the embedding according to the ground truth to facilitate the analysis; in real applications, this dimensionality could be estimated empirically during scale estimation (cf. Künstle et al., 2022a).

SOE loss:

$$\arg\min_{\hat{\boldsymbol{x}}} \sum_{(i,j,k)\in T} \max\left[0, \mathrm{d}(\hat{\boldsymbol{x}}_j, \hat{\boldsymbol{x}}_i) - \mathrm{d}(\hat{\boldsymbol{x}}_i, \hat{\boldsymbol{x}}_k) + 1\right]^2$$

Metrics

We calculate the validation triplet error from each estimated scale as a quality measure (e.g., Haghiri et al., 2020; Roads & Mozer, 2019) and compare it with the recovery error to ground truth. The triplet error is calculated by counting triplets that are inconsistent with the scale. While the error on *training* triplets, i.e. the triplets used to run the ordinal embedding algorithm, can be used as a measure of fit, the quality is typically measured by a *validation* or *test* error. For this purpose, we calculate the test error on independently sampled test data from the same ground truth; in applications without access to ground truth, the test error is typically approximated with cross-validation (cf. Haghiri et al., 2020).

Triplet error:

$$\frac{1}{|T|} \sum_{(i,j,k)\in T} \begin{cases} 0 & \text{if } d(\hat{\boldsymbol{x}}_i, \hat{\boldsymbol{x}}_j) < d(\hat{\boldsymbol{x}}_i, \hat{\boldsymbol{x}}_k) \\ 1 & \text{otherwise.} \end{cases}$$

When calculating the difference between the validation or test error and the training error, we obtain the *train-test gap*. A small gap typically indicates a lack of overfitting.

The *recovery error* measures the Euclidean distances between the scale estimate and ground truth. We take this error under optimal similarity transformation of the scale by aligning with ground truth through a so-called Procrustes analysis (Gower, 1975).

Recovery error:

$$\sqrt{\sum_{n=1}^{N} \sum_{d=1}^{D} (\boldsymbol{\psi}_{nd} - \hat{\boldsymbol{x}}_{nd})^2}$$

4.2.2 *Comparison of metrics to ground-truth*

We vary the number of triplets and the noise intensity with the factors $\lambda \in \{\frac{1}{4}, \frac{1}{2}, 1, 2, 4\}$ and $\omega \in \{0, 1, 2, 3, 4\}$. λ and ω were chosen to simulate a broad but still realistic range of experiments (cf. Künstle et al., 2022a); in practice, these values could be influenced consciously by the design of the experiment, for example, by collecting additional trials or improving the stimuli. Compared to λ and ω, the noise distribution in the simulation or the embedding algorithm's loss function typically have a minor influence (Maloney & Yang, 2003; Roads & Mozer, 2019) and are kept constant here. We get an impression of the stability of the estimates by generating ten random data sets for each $\lambda - \omega$ combination, which we embed and evaluate with the reconstruction and triplet error (on training data and independently sampled test datasets).

In Figure 4.1, both metrics show the expected progression that more

data and less noise reduce the error. But that's almost where the similarities end. The error decrease with more data is almost linear in the reconstruction error (with our parameterization) and approaches zero for $\lambda \geq 1$. This dataset size corresponds to the theoretically grounded $DN \log_2 N$ rule-of-thumb to choose the number of trials in a triplet experiment (Haghiri et al., 2020; Jain et al., 2016). The noise level increases the average and spread of the reconstruction error but can be compensated by larger dataset sizes (i.e. for $\lambda \in \{2, 4\}$ and $\omega \in \{0, 1\}$). Neither error variation nor compensation is visible in the triplet error, which casts doubt on its ability to predict the reconstruction error.

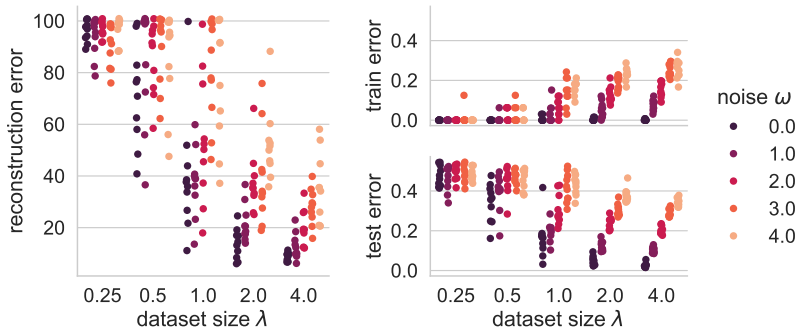


Figure 4.1: The reconstruction and triplet error on training and test datasets generated with the lightness simulation.

The difference of the triplet error between training and test data set seems to be visually much more related to the noise (already observed in Künstle et al., 2022a). In fact, the mean train-test gap correlates highly with the mean reconstruction error (Figure 4.2), both in the mean and variation ($p < .001, r = .97$ and $p < .001, r = .80$;).

This result implies that we can predict the reconstruction error—unknown but highly relevant in real experiments—from the train-test gap, which can be approximated empirically. The train-test gap thus allows the experimenter to make statements about the reliability of the estimated scale from almost random (gap: 0.5) to perfect reconstruction (gap: 0). Because the gap is linear to the actual deviation in the scale space, it is particularly straightforward to interpret. It helps in decisions such as whether collecting more trials could benefit the scale estimate.

The train-test gap is a summary metric on the whole scale, which can only be used to interpret the overall potential deviation from ground truth, but not its distribution. However, to interpret a scale, the local variation in scale space is at least crucial; we look at this in the following section.

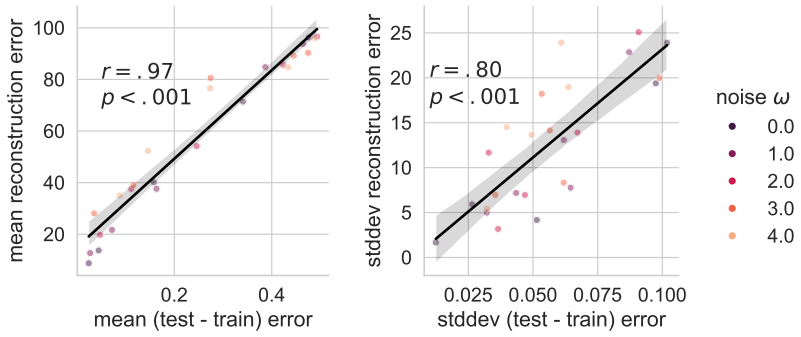


Figure 4.2: The reconstruction error is highly correlated with the difference between the train and test triplet errors. Scatters show the metrics for embeddings of datasets with varying size and noise (see text); the mean and standard deviation are calculated over 11 repetitions each.

4.3 *Stability estimation procedures*

Visual interpretations of psychophysical scales, especially those indirectly estimated with triplets, would benefit from visualizations of putative variation in the form of error bars, confidence intervals, and confidence regions to prevent misinterpretation.

In the past, methods based on resampling or probabilistic models were primarily used to quantify variability. Here, we want to evaluate and discuss both approaches using the simulations presented in the previous section.

4.3.1 *Resampling perspective*

Determining the variability of a model using resampling techniques, in particular *bootstrapping* (Efron, 1987), always follows the same pattern: Several data sets are sampled from a single data set, on each of which the model is fitted and the parameter variability is then statistically evaluated (e.g., Tibshirani, 1994; Wichmann & Hill, 2001b).

Error bars of scales with MLDS are generated this way, with bootstrapping (Knoblauch & Maloney, 2012a; Maloney & Yang, 2003). Starting with an MLDS fit to the original data set, the data sets are resampled with noise. In contrast, Lohaus et al. (2019) suggests generating the data sets by simply subsampling the triplets (not the stimuli). They showed that the overall variation follows the expected trend, decreasing with the size of the dataset and increasing with the noise.

What all these bootstrapping approaches have in common, however, is that the scale samples must first be aligned before variability can be determined locally.

The alignment problem

TRIPLET COMPARISONS do not uniquely define a scale's translation, rotation, scale, or flip; the scale samples must be aligned respectively.

A common alignment procedure is to select a stimulus as the origin and to fix the scaling or orientation of another one, for example, using a noise parameter (in MLDS fitted with `glm`, see Knoblauch & Maloney, 2012a) or simply to one (Knoblauch & Maloney, 2012a; Lohaus et al., 2019). This procedure, which we call $0-1$ *alignment*, is usually only used for 1D scales but could hypothetically be extended by selecting a stimulus for each dimension that defines the positive direction[3].

In contrast to $0-1$ alignments, the *Procrustes alignment* treats all stimuli equally by minimizing the (Euclidean) distance across scales (cf. Lohaus et al., 2019; Sauer et al., 2024). The minimal transformation to align two scales can be found analytically. Multiple scales, however, must be aligned with the *generalized Procrustes analysis* (Gower, 1975). This analysis iteratively adjusts the scales until they no longer change. In each iteration, the scales are individually aligned to the common pointwise mean using the standard Procrustes alignment.

[3] This multidimensional variant of $0-1$ alignment is inspired by the prior of a Bayesian MDS model of Gronau and Lee (2020). This prior was developed to to ensure identifiability in a multidimensional representation.

BOTH alignment methods bias the distribution of the variability within the scales differently. $0-1$ alignment inevitably leads to the variability estimate being zero for D or $D+1$ stimuli. While it may be more of a philosophical consideration whether there is no variability of "zero loudness" or "zero brightness", the case is much more apparent for most other stimuli without such a natural origin. No hue is a "zero hue", and no animal is a "zero animal", so defining one of the stimuli seems arbitrary. Even more concerning is the observation in our color simulation, suggesting that the $0-1$ alignment of multiple scale samples exaggerates the variability of the non-fixed points and distorts it depending on the choice of orientation stimuli (see Figure 4.3).

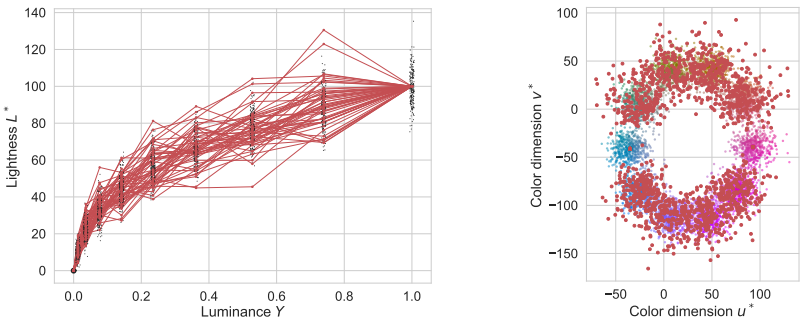


Figure 4.3: The $0-1$ alignment of the simulated lightness and color scales. The variability of the scale (red dots) is zero for two stimuli but amplified for the others in contrast to the ground truth samples (black/-colored dots).

The Procrustes alignment, on the other hand, "blurs" all local vari-

ability. By definition, the quadratic differences of all stimuli and all dimensions are minimized in equal parts. This means that all scale variances are approximately the same and that large or small variability of individual stimuli or dimensions is not recognizable. We can imagine a thought experiment where normally sighted and colorblind people are asked to judge triplets of objects based on shape and color. We might expect the colorblind to respond with significantly more variability based on color but not shape than the normally sighted participants. However, Procrustes aligned bootstrapped scales would diminish those differences in the scale dimensions. Figure 4.4 demonstrates the effect with samples from our lightness simulation: The variability of all aligned stimuli is about the same. However, the noise of the responses increases strongly with the (simulated) luminance.

We must conclude that these forms of alignment always distort local variability and can only be useful as a global metric. This alignment problem leads us to look for a solution beyond bootstrapping for a local variability measure.

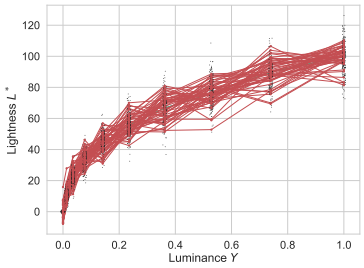


Figure 4.4: The Procrustes alignment of simulated lightness scales in red; ground truth samples are shown by transparent grey scatters.

4.3.2 *Probabilistic perspective*

Here, we propose a change of perspective to assess the scale's stability: Instead of bootstrapping new scales, we fall back on a single fitted scale and "wiggle" it. If the triplets firmly determine the scale, it should have little wiggle room without dropping the goodness-of-fit. We use probabilistic modeling methods to implement "wiggling" efficiently and validate the obtained measures using the simulations from the previous section.

Evidence of triplets

There is a geometric perspective on how triplets determine a scale—or just give it room to "wiggle" (cf. Jamieson & Nowak, 2011).

We can think of each triplet (i, j, k) as a bisection of the perceptual space of one-half closer to $\hat{x}_j$ and the other closer to $\hat{x}_k$. If the triplet response is correct, then the half-space of $\hat{x}_j$ must contain $\hat{x}_i$. The half-spaces of further triplets overlap and restrict the putative region where $\hat{x}_i$ is located (Figure 4.5); in the limit, this region converges to a point that fits the triplet responses.

Since triplet responses can be false or contradictory, we might adopt a stochastic description and say the triplets increase the evidence of a region. The evidence of conflicting responses can cancel each other out and thus increase the putative region.

Mathematically, we can describe this evidence for the embedded location of a stimulus as $P_{\hat{x}_i}$, the product of the likelihood of all triplets

P^i_{jk} involving the stimulus[4].

This likelihood must be a function that describes the correctness of the triplet for the given scale (Tamuz et al., 2011) and can be found in the loss function of various ordinal embedding algorithms (Lohaus et al., 2019).

The evidence in the perceptual space, as visualized by the red area in Figure 4.5, can then be determined by shifting $\hat{x_i}$ ("wiggling") and fixing the other points (Tamuz et al., 2011). However, since the other stimuli are not determined to a point, we would have to wiggle them simultaneously. This quickly becomes computationally infeasible, and we must look for a solution using probabilistic inference methods.

Our stability model

We define a probabilistic model that allows us to determine the evidence for alternative scale locations efficiently.

An MCMC sampler generates scale locations proportional to their posterior probability. The posterior is specified by the likelihood of all triplets involved and a prior (Figure 4.6).

We calculate the likelihood of a triplet with the fraction of distances in the scale[5] as introduced in the loss of the CKL embedding algorithm (Tamuz et al., 2011). This likelihood has two advantages: Firstly, it is size-independent and can, therefore, do without standardization or additional parameters. Secondly, we can interpret the likelihood so that the uncertainty of a triplet decision increases with perceived dissimilarity. This is a psychologically plausible inductive bias for many stimuli; for example, we expect variability in kilometers when judging distances of (Paris, Berlin, Madrid), but only meters for (the Eiffel Tower, Arc de Triomphe, Notre-Dame de Paris).

The prior specifies the initial probability of scale samples and has the particular task of keeping the model identifiable by constraining similarity transformations (Gronau & Lee, 2020; Lohaus et al., 2019; Roads & Mozer, 2019). Variational inference methods would be an alternative approach to enforce identifiability (Muttenthaler et al., 2022; Roads & Love, 2021). However, we will not use variational inference in this work, since it approximates the posterior with a Gaussian distribution and thus symmetrically, which is not necessarily appropriate in the psychological space.

We center our prior distribution on the scale whose variability we want to determine, varying the scaling on the distance to the neighboring stimulus. In the following, we use a Student-t distribution with D degrees of freedom and a standard deviation with half the distance to the $D+1$-nearest neighbor. We use nearest neighbors to adaptively adjust the prior width based on the scale's density. This is important

[4] Evidence of stimulus l at x_l:

$$P_{\hat{x}_l} = \prod_{(i,j,k)\in T} \begin{cases} P^i_{jk}, & \text{if } l \in i,j,k \\ 1, & \text{otherwise.} \end{cases}$$

Note that Tamuz et al. (2011) use a similar formula but restrict the if-condition to $l = i$.

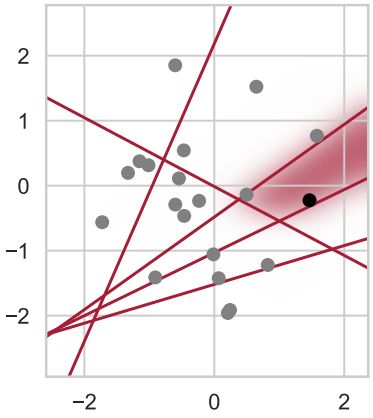


Figure 4.5: A 2D scale with a highlighted stimulus (black), whose location should be determined by triplets. Five triplets partition the space (red lines) and mark down the region where the highlighted stimulus is likely located (red shade).

[5] CKL triplet likelihood (Tamuz et al., 2011):

$$P^i_{jk} = \frac{d(x_i, x_k)^2}{d(x_i, x_j)^2 + d(x_i, x_k)^2}$$

because we assume that the maximum likelihood estimate of the scale, i.e. the mean of the prior, is basically in the right place. The prior enforces that the stimuli cannot completely change location or order in the scale but only "wiggle" slightly in their neighborhood. As with any informative prior, it is advisable to validate its impact on scientific results by varying the prior. We demonstrate this in one of the following application examples (subsection 4.4.1).

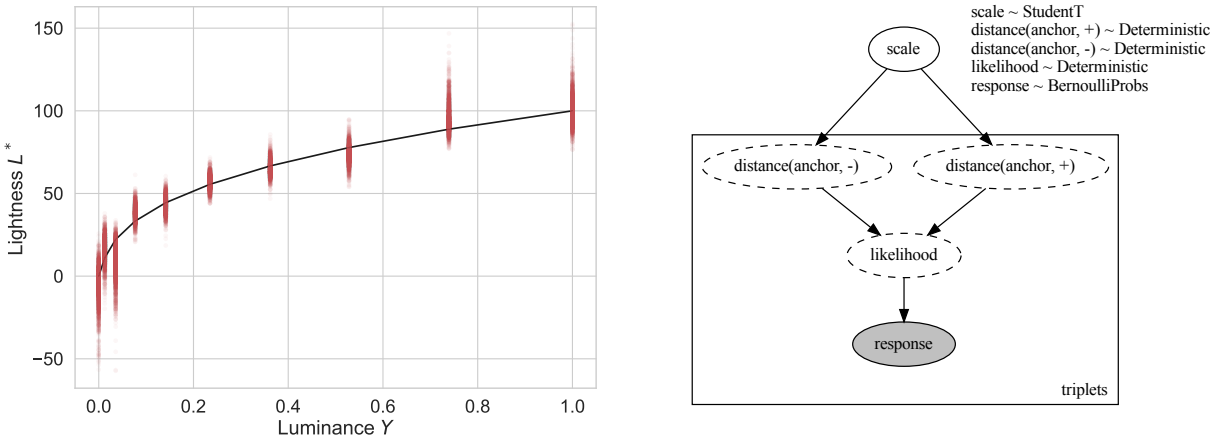


Figure 4.6: Probabilistic model for estimating the variability of a psychophysical scale. The MCMC samples (red, left plot) indicate the variability and were generated according to the model graph (right plot).

With the help of modern inference methods, the probabilistic programming library `numpyro` (Bingham et al., 2019; Phan et al., 2019), the No-U-Turn Sampler (NUTS; M. D. Hoffman & Gelman, 2014), as well as automatic reparametrization (Gorinova et al., 2020), the variability on psychophysical scales (like the ones simulated) can be estimated in seconds. This inference is often noticeably faster than estimating hundreds of scales from bootstrap samples.

If we fit the probabilistic model to the triplets of the lightness simulation and plot the Monte Carlo samples in Figure 4.6, we see a distribution of scale values. This scale distribution is very different from the bootstrapped samples in Figure 4.3 and Figure 4.4. In contrast to the Procrustes aligned samples, the variability here increases with luminance; this increase may reflect the increase in noise in the simulated ground truth. In addition, the variability is particularly high for the highest and lowest luminance stimuli, where 0-1 aligned bootstrap samples showed zero variability. These edge values may be less determined by their neighbors, plausibly explaining the reduced stability shown here.

As a quantitative evaluation of these variability measures, we repeat their estimation 100 times on the lightness and color simulations, varying $\lambda \in 1, 2, 4$ and $\omega \in 0, 1, 2$. For each estimate, we form so-called high posterior density intervals (per dimension) with a probability mass $p \in 0.5, 0.9, 0.99$ and check whether the ground truth points lie within the intervals. The hit rate, called coverage, should match the corresponding p as closely as possible. Figure 4.7 shows that for $\lambda \geq 2$ the coverage largely matches the probability mass (horizontal line). At $\lambda = 1$, the scale estimators on which our prior is based still deviate strongly from ground truth, which makes reliable variance estimation difficult. The intervals, which are a rough approximation of the 2D variability, can also explain the coverage overestimations in the color simulation.

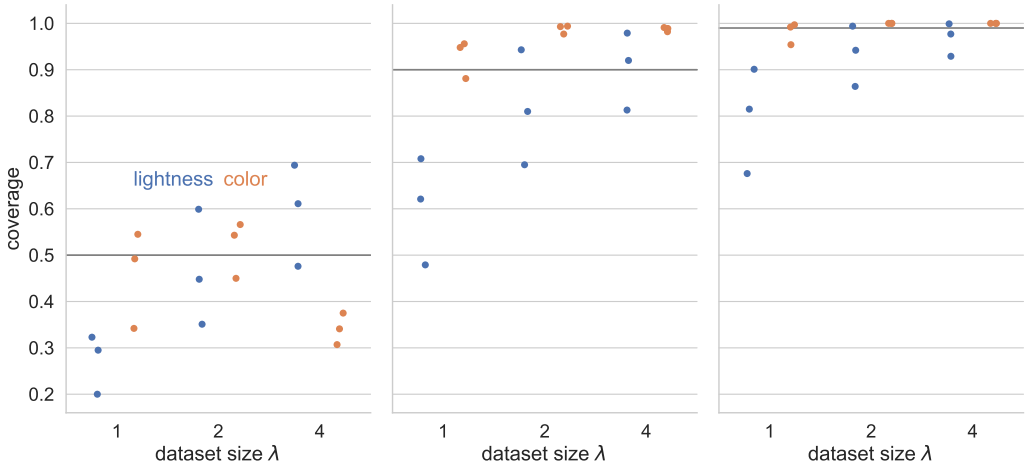


Figure 4.7: The uncertainty coverage in repeated simulations of lightness and color scales with varying numbers of triplets and noise levels. The coverage should match the probability mass, indicated as horizontal lines.

While simulation experiments are essential to validate a new methodology qualitatively and quantitatively, they can never fully reproduce reality. In the following section, we apply it to three behavioral datasets.

4.4 *Quality and stability in behavioral datasets*

Here, we show the potential of stability determination on three diverse data sets from real experiments. In the first two datasets we use estimation to determine the scale's sensitivity and the subject's reliability, respectively. In the third example, we use variability estimation to gain insight into a dataset that is too large and high-dimensional for classical chi-by-eye.

4.4.1 *Visualizing the variability*

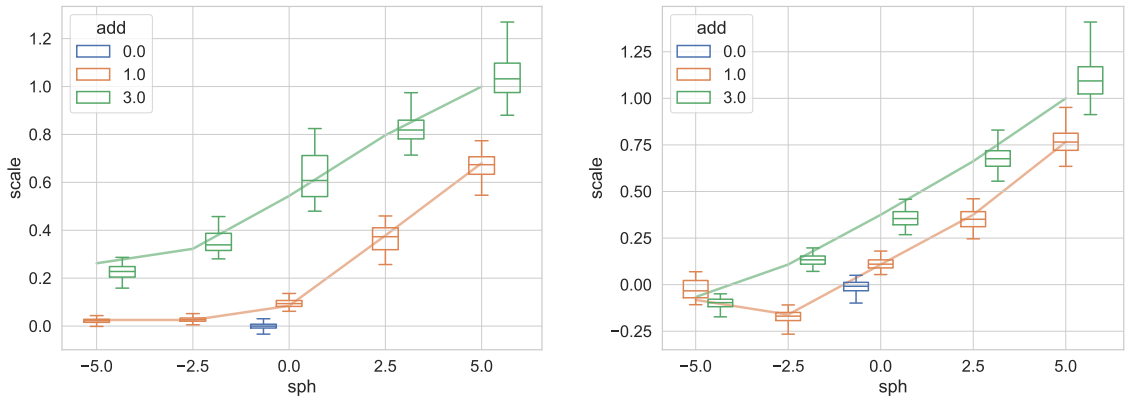
When a psychophysical scale is visualized, interesting structures might not be interpreted as the effect of the observer's behavior but merely as an artifact of the modeling. The scale's variability must also be visualized to distinguish effect from artifact.

In this example, we use triplets from the experiment described in detail in Sauer et al. (2024). In this experiment, observers compared the similarity of simulated distortions caused by progressive spectacles with different spherical (*Sph*) and addition (*Add*) corrections.

The embedding dimensionality was fixed to $1D$ based on the cross-validating triplet error (see chapter 5 for details); the gap between training and test error is very small for most subjects. According to the results in the previous section, this indicates that the $1D$ scale has a high quality in reconstructing perception.

A noticeable difference between the observers' scales is the influence of *Add*. Scales differ in their effect of the *Add* level, indicating that observers perceive the addition power with different intensities. But is this difference real, or is it just an artifact of the variability of the embedding?

In Figure 4.8, we show the scales of two observers together with a boxplot of Monte Carlo samples of the probabilistic model described in the previous section.



(a) Observer 1

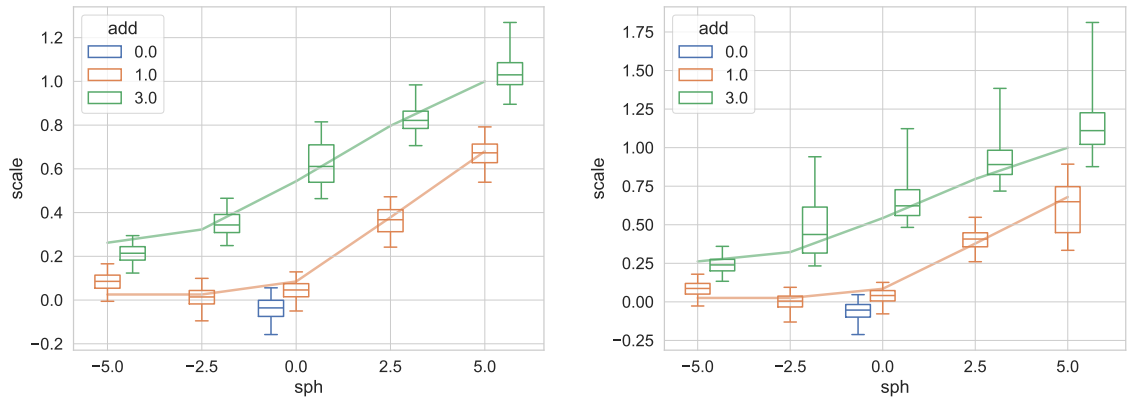
(b) Observer 5

Figure 4.8: The scale estimate (line) and variability (boxplot) for two observers. The boxplots show median and quartiles, while the whiskers show 2.5 and 97.5 percentiles of the Monte Carlo samples.

The boxplots at different *Add* levels do not overlap for negative *Sph* in observer 1's scale—we can, therefore, assume that they perceived distortions at *Add* levels differently. For observer 5, on the other hand, the whiskers and boxes for different *Add* levels overlap at *Sph* −5dpt and 5dpt and nearly overlap at others *Sph* values. This overlap could indicate that observer 5 was not able to reliably distinguish the distortions based on *Add* levels for strong spheric corrections. Since *Sph*

and *Add* affect stimulus distortions in different areas of the visual field, this difference in sensitivity to addition correction between observers could indicate that the observers used different strategies to judge the triplets.

An alternative explanation for the difference in addition sensitivity between observers could be that our probabilistic model overestimates the stability of observer 1 due to the small scale differences for negative *Sph* values. The default prior to the probabilistic model is based on the distance to the two nearest neighboring points. The almost constant scale for negative *Sph* values results in very narrow priors. To validate that the scale's gap between different *Add* values is not just an artifact of the narrow prior, we rerun the analysis of observer 1 with a constant prior of the standard deviation of 0.1 for all stimuli and an adaptive prior based on the four nearest neighbor distances. Although some boxplots in Figure 4.9 that were previously separated now overlap slightly, the differences between the *Add* levels are still clearly recognizable, confirming our observations.



(a) Prior fixed

(b) Prior 4-NN

Figure 4.9: The variability samples were generated with wide priors, either standard deviation 0.1 for all stimuli or half the distance to the fourth nearest neighboring point.

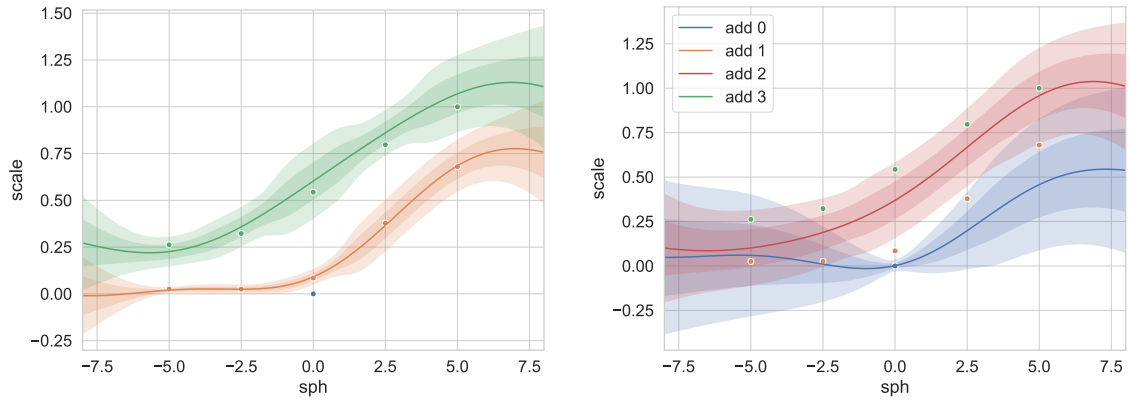
4.4.2 *Interpolating the stimulus intensities*

The measured lenses only represent a limited range of possible spectacles. Therefore, by interpolating the scale, it would be interesting to show the perceived distortion of other *Sph* and *Add* values. In this section, we would like to show an interpolation of the scale based on our probabilistic model that accounts for the uncertainty of the scale estimation and the interpolation itself.

First, we pick some of the Monte Carlo samples from our probabilistic model and fit a Gaussian process regressor (GPR). The Gaussian process is trained to predict the scale value from the features *Sph* and *Add*. The Monte Carlo samples prototypically represent the vari-

ability at the measured scale values. Both the mean scale value and its variability are interpolated between stimuli by the GP's kernel. In this example, it is important to ensure that the noise of different stimuli can vary in size. Therefore, we combined an RBF kernel with a heteroscedastic noise kernel that interpolates the variability between "prototype datapoints" (i.e., the measured stimuli). We implemented the GPR with `scikit-learn` (Pedregosa et al., 2011) in combination with the noise kernels by the `gp_extra` library (Metzen, 2016).

We interpolate the scale of observer 1 by randomly selecting 20 of the Monte Carlo samples from the probabilistic model (see subsection 4.4.1) and fitting the GPR. Figure 4.10 shows the predicted scale by their mean and standard deviations for $Sph \in [-7.5, 7.5]$ and $Add \in \{0, 1, 2, 3\}$. The mean value (lines) closely follows the original scale values (dots) and shows a smooth transition in between. The standard deviations also show a smooth transition between the stimuli of the original scale. The Gaussian assumption of the GPR results in symmetric variability, deviating from the asymmetric distribution of the probabilistic model in Figure 4.8. The mean value flattens out for *Sph* values outside the measured range, and the standard deviation widens considerably. Predictions for lenses with *Add* 2dpt, which did not occur at all in the experiment, have a slightly larger standard deviation but otherwise follow a curve between predictions of *Add* 1dpt and 3dpt. These extrapolations have the potential to use the scale as a basis for a more general perceptual model in the form of the GPR and, thus, in downstream applications, for example, to predict the perceived distortion of spectacles, even if their lenses were not part of the original experiment.



(a) Interpolate & extrapolate *Sph*

(b) Interpolate *Add*

Figure 4.10: The interpolated and extrapolated scale of observer 1 as a prediction of a heteroscedastic Gaussian Process. The lines show the mean (lines) prediction, while the shaded areas show the first and second standard deviations. The dots show the original scale values.

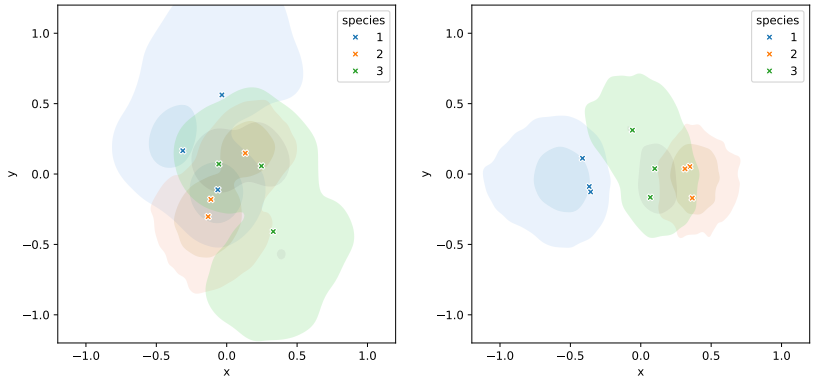
4.4.3 *Interpreting stability as response consistency*

Here, we present an example showing that modeling embedding stability allows us to draw conclusions about our scaling model and the observers' response behavior. The application uses data from an experiment where observers of different expertise rated triplet comparisons between mushroom images[6]. The core idea of the experiment is that for a layperson some mushroom species appear very similar but an expert must distinguish them reliably (e.g. to detect toxic mushrooms); thus, internal representations of laypeople and experts should be different. Both layperson and expert judged triplets sampled from nine mushroom images. Three images each of chanterelles (*cantharellus*), false chanterelles (*hygrophoropsis aurantiaca*), and jack-o'-lantern mushrooms (*omphalotus olearius*) were presented. The observers were instructed to rate the images according to their similarity.

[6] We thank Lukas Huber and Raphael Wittwer (University of Bern) for making this dataset available to us.

We estimate scales from the triplets with SOE (Terada & von Luxburg, 2014) in 2D to simplify the visualization and approximate the scale variability with the probabilistic model introduced previously in this chapter. For each species, we estimate the 2D density of the Monte Carlo samples with a Kernel Density Estimator (KDE).

Figure 4.11 shows the scales of two observers with regions of 65% and 95% probability mass. From the point estimates alone, we can deduce that the layman confuses the mushroom species, but the expert can distinguish them reliably. The layperson could most likely not classify the mushrooms by species and responded inconsistently. These inconsistent responses lead to large and overlapping regions of putative scale representations.



(a) Layperson (b) Expert

Figure 4.11: The two-dimensional embedding of two observers, judging triplets of mushroom images. The shaded area shows one and two standard deviations of a density esti than the ground-truth variability (black/colored pointsimate of our probabilistic model.

For the expert, only the regions of species 2 and 3 overlap slightly; species 1 is clearly separated. This suggests that the expert was able to reliably distinguish species 1 from the others, but that the other

two species were not always strictly distinguished. This is plausible because only species 1 is an edible mushroom. Mushroom experts are trained to distinguish edible from non-edible mushrooms.

4.4.4 *Analyzing high-dimensional structures*

The last application example deals with the analysis of large embeddings. In an online study, Hebart et al. (2020) showed observers triplets of 1,854 object images from the Things database (Hebart et al., 2019) and asked them to choose the *odd-one-out*. From the answers, they estimated a scale with 49 sparse dimensions that should encode individual perceived properties with positive numbers. Here we use our probabilistic model to analyze the stability of the object representations. Consistent with the sparse embedding, we measure distance using the city block metric instead of the Euclidean metric.

To analyze stability, we use our probabilistic model with the publicly available pre-trained scale and a publicly available dataset of test trials, i.e. trials that were not used to compute the scale. In order to speed up the estimation significantly, we simplify the probabilistic model. We estimated the posterior probability independently for each stimulus, assuming that the representation of the others is fixed. The posterior is inferred with a uniform prior, centered around the scale estimate and scaled by the city block distance to the 10-nearest neighbors. Similar simplified inference techniques are used by Tamuz et al. (2011) to determine the most informative trials efficiently. With this simplified model, the variability of all data points can be inferred in about 6 minutes on a regular laptop.

We use the posterior density around a data point to express its stability. The stability in this dataset is related to the inverse density of the embedded points, i.e., the average distance of a point to its 100 neighbors (Figure 4.12). This corresponds to our interpretation of stability: Points are constraint by their neighborhood.

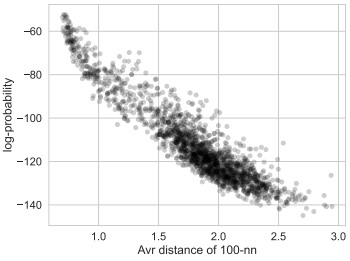
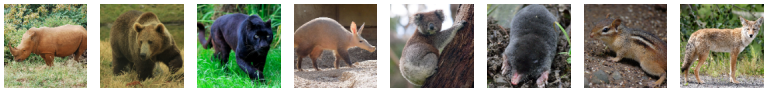


Figure 4.12: The variability correlates with the inverse density of the representation.

Figure 4.13 shows the top 8 most stable and variable points. Here, we also see substantial differences in the content of the images. The most stable embedded objects are animal images, many of which are in the dataset, and it is probably easy for observers to evaluate them as similar. The most variable representations, on the other hand, show objects that cannot necessarily be identified at first glance. We might speculate that observers could not identify those objects reliably and thus judged them more randomly than other stimuli.

By quantifying this variability, we can evaluate which stimuli in the scale we can trust–and which we cannot–even without visualizing the entire scale. These differences in variability between stimuli open up a new aspect in the interpretation of a scale: In addition to dimensions

and distances between stimuli, variability can also be interpreted as a measure of judgment uncertainty.



(a) Most stable

(b) Most variable

Figure 4.13: The top-8 most variable and most stable represented images. The images are provided by the Things database for research purposes (Hebart et al., 2019).

4.5 *Discussion*

In previous sections we have examined and practically applied methods to determine the quality and stability of psychophysical scales determined with ordinal embedding algorithms.

In simulation experiments the validation triplet error, which is typically used as a quality measure (cf. Haghiri et al., 2020), allows only insufficient conclusions to be drawn about the "correctness" of the scale estimators, i.e., the recovery error to a ground truth unknown in real experiments. It is possible to obtain scales with low recovery error even with a high triplet error. Therefore, the validation error may well be a measure of the quality of the response prediction but not of the embedding. The train-test gap, i.e., the difference between the training and validation triplet errors, was a much more interpretable metric of embedding quality in our simulations.

With the ability to confirm the quality of a scale estimate, we have created the necessary conditions to estimate the scale's variability. Variability estimates depend on the assumption that the estimated scale is close enough to the ground truth (cf. the *bootstrap bridge assumption*; Wichmann & Hill, 2001a), which must be tested empirically via the train-test gap being almost zero.

First, we looked at bootstrapping methods and found that, in general, there is a risk of biasing variance estimators locally. The alignment problem states that the bootstrapped sampled scales must first be matched with similarity transformations and—depending on the alignment procedure—variance information is shifted differently. Therefore, we introduced a probabilistic model to estimate variability based on the intuition that consistent triplets would restrict the freedom of movement of the scale.

The model provides samples of scales distributed according to their

probability of predicting the triplet responses. These samples can be aggregated to variability metrics and visualized as credible intervals or regions. In fact, such regions could not only be interpreted as model variability, but they are also congruent with the concept of psychological spaces. For example, Shepard (1987) mentioned that an object is part of a class, which "corresponds to some region in the individual's psychological space".

From this perspective, we consider the potential of combining bootstrapped scales and their probabilistic variability estimates. The bootstrapped samples, represented as single points or lines and Procrustes aligned, can represent alternative "discrete" trajectories of the estimate.

We refer to our variability model in the text as probabilistic but not as Bayesian, since we use probabilistic programming techniques but do not strictly apply Bayesian statistics. For example, in most cases we use the identical triplets to configure the prior and to infer the posterior. Our interpretation is that we the probabilistic inference as a tool to determine the fit of "wiggling" scale alternatives. However, if enough trials can be collected it might be more correct to use a separate dataset set during inference, like we do in the last application example.

This problem vanishes if scaling and variability estimation are not separate steps, as in this work, but would be combined. Such approaches of probabilistic ordinal embedding methods (Lohaus et al., 2019; Muttenthaler et al., 2022; Roads & Love, 2021; Roads & Mozer, 2019) seem very tempting and should, in our opinion, be further investigated and, above all, validated. In this work, we have explicitly taken an algorithm-agnostic approach, because in this way we are able to better support the psychophysicist. They can continue to use the embedding algorithms they are familiar with, such as MLDS, and perform the variance estimation separately. Furthermore, as shown in the last application example, the algorithm-agnostic approach allows us to investigate embeddings without having access to all training data.

There are many more applications for quality and stability estimates. In addition to the possibilities shown in our application examples of evaluating the sensitivity of the scale estimate better and thus preventing false scientific conclusions, they can also be useful in downstream tasks.

Adaptive sampling (also known as active learning) methods, for example, are based on collecting only informative trials and require robust uncertainty estimates. Since we have obtained some methods from this literature for this work, we would be pleased if our variance estimation could be helpful in this application.

On the other hand, there have been attempts for some time to develop a link between scaling and threshold methods (e.g., Aguilar et al., 2017; Devinck & Knoblauch, 2012). For this purpose, it is crucial to obtain accurate variance estimates of the scale to predict just-noticeable differences. Here, we would rely on variability estimates with separate test triplets since with a good fit and comparable to the *noise ceiling* of the triplet error in our simulation experiments, this could be interpreted as perceptual or response noise.

TOGETHER with the toolbox from chapter 2 and the dimensionality testing procedure from chapter 3, the most important methodological gaps for the successful use of ordinal embedding algorithms in psychophysics have been closed. We will demonstrate the scaling pipeline on a experiment in the next chapter.

5 Psychophysical scale of spectacle lens distortions

The great painters are incomparable draftsmen. They also know how to mix their own paint, grind it, put in the fixative; no task is too small to be worthy of their attention.
— Twyla Tharp
(Tharp, 2006)

The strength of ordinal embedding methods in measuring perception is demonstrated in particular by the intuitive triplet task, the comparatively small number of trials, and the robustness against subject errors. These properties enable the study presented in this chapter to measure the distortion of varifocals. The distortions investigated are difficult to compare and can only be recognized in dynamic virtual reality stimuli. Nevertheless, we obtain the most comprehensive scales of PAL distortions to date and thus lay the foundation for perception-based spectacle treatments.

Both the content and form of this chapter equal our published article[1], licensed under CC BY 4.0, and the corresponding preprint (Sauer et al., 2023a). Yannick Sauer and I contributed equally to this study and wrote the majority of the article. I implemented the ordinal embedding analysis while Yannick Sauer implemented the VR experiment and collected the data. All authors contributed to the scientific idea and the writing.

[1] Sauer, Y., Künstle, D.-E., Wichmann, F. A., & Wahl, S. (2024). An objective measurement approach to quantify the perceived distortions of spectacle lenses. *Scientific Reports*, *14*(1), 3967

Abstract

The eye's natural aging influences our ability to focus on close objects. Without optical correction, all adults will suffer from blurry close vision starting in their 40s. In effect, different optical corrections are necessary for near and far vision. Current state-of-the-art glasses offer a gradual change of correction across the field of view for any distance—using Progressive Addition Lenses (PALs). However, an inevitable side effect of PALs is geometric distortion, which causes the *swim effect*, a phenomenon of unstable perception of the environment leading to discomfort for many wearers. Unfortunately, little is known about the relationship between lens distortions and their perceptual effects, that is, between the complex physical distortions on the one hand and their subjective severity on the other.

We show that perceived distortion can be measured as a psychophysical scaling function using a VR experiment with accurately simulated PAL distortions. Despite the multi-dimensional space of physical distortions, the measured perception is well represented as a 1D scaling function; distortions are perceived less with negative far correction, suggesting an advantage for short-sighted people.

Beyond that, our results successfully demonstrate that psychophysical scaling with ordinal embedding methods can investigate complex perceptual phenomena like lens distortions that affect geometry, stereo,

and motion perception. Our approach provides a new perspective on lens design based on modeling visual processing that could be applied beyond distortions. We anticipate that future PAL designs could be improved using our method to minimize subjectively discomforting distortions rather than merely optimizing physical parameters.

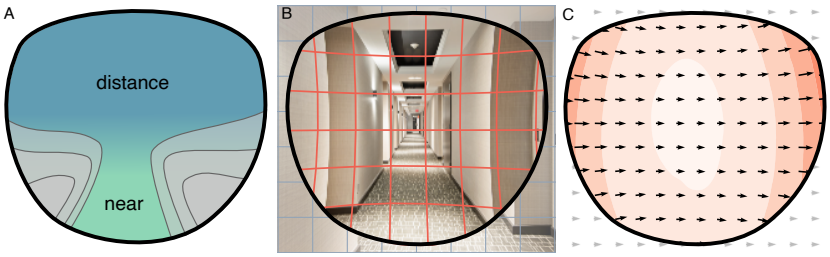


Figure 5.1: Progressive Addition Lenses (PALs). A) The upper lens area is designed for far vision, with an optical power fitting the far refraction of the wearer. The optical power increases vertically towards the near area, which offers additional power for focusing close objects. The gradient in power will always lead to lateral astigmatism. B) Optical distortions of PALs change the size and orientation of objects. Vertical and horizontal edges in a typical indoor environment appear curved. C) Perceived motion during a horizontal head movement. The retinal motion pattern—optic flow—is altered by optical distortions. Points in the visual field move along curved trajectories instead of straight lines. Lens distortions increase towards the periphery leading to an increase in optic flow speed (illustrated by the heat map in the background). The unnatural distorted optic flow pattern can be perceived as an unstable movement of the environment (swim effect).

5.1 *Introduction*

THE NATURAL DECLINE in the eye's accommodative capabilities makes it progressively harder to focus on close objects. This natural aging process results in the eye's lens becoming less flexible, losing its ability to change shape. Approximately 2 billion people worldwide suffer from this condition, termed *presbyopia* (Charman, 2014; Fricke et al., 2018); without optical correction, all adults would experience blurry close vision starting in their 40s.

Presbyopes can use reading glasses for sharp near vision, which have to be taken off for looking at distant objects—in the case of preexisting corrections, this requires constant switching between two pairs of glasses. The convenient solution combines both lenses: the near correction at the bottom and the far correction at the top of each lens (Letocha, 1990). Those so-called bifocal lenses, however, cannot offer correction for intermediate distances and show a visible, distracting border at the transition between both lens areas. This edge in the lens is also considered to create a stigma of glasses for "old people". An obvious refinement is a smooth transition between the near and far areas by gradually changing the curvature of the lens surface. These Progressive Addition Lenses (PALs) became available in the second half of the 20th century through technical advances in manufacturing technologies (Pope, 2000; Sullivan & Fowler, 1988). Today, PALs are the state-of-the-art lens in presbyopia correction.

EVEN THOUGH there has been great progress in improving PALs, the gradual increase in optical correction between far and near areas will always lead to unwanted optical errors, so-called aberrations. This causes blur or degradation of sharpness in some areas of the lens.

Another aberration is geometric distortion, a variation in the magnification across the visual field, leading straight lines to appear curved when looking through the lens (Figure 5.1). Sadly, for PALs, it is physically impossible to reduce aberrations to zero, as stated by the Minkwitz theorem (Meister & Fisher, 2008; Minkwitz, 1963; Sheedy et al., 2005). What lens designers and manufacturers can do, however, is to *change the distribution of aberrations* in the field of view—attempting to find subjectively more benign patterns of aberrations across the visual field.

This flexibility in shifting the aberrations to different areas in PAL design is used to develop specific PALs for tasks like driving or office work. Such PALs show reduced blur in task-relevant areas, inevitably accompanied by increased blur in other areas, however.

How to optimize the design for distortions *in general* is an open question, however, because their influence on visual perception is poorly understood. The swim effect, a phenomenon of unnatural or unstable perception of the environment during head or eye movements, causing instability, dizziness, tripping, and nausea (Alvarez et al., 2009; Johnson et al., 2007; Meister & Fisher, 2008; Sauer, Scherff, et al., 2022), is at least partly caused by geometric distortions. It is unclear how those effects scale with the physical distortion of the lens and its distribution across the visual field.

Understanding and quantifying the influence of distortions on human perception can——in future applications—enable PALs with reduced distortion-induced discomfort. To our knowledge, this study presents the first rigorous measurements of perceived distortions of PALs.

The main influence on PAL distortions is the optical correction in the far and near area, the optical power measured in diopters. The correction in the far area, the spherical power *Sph*, can be positive or negative (correction for hyperopia or myopia, respectively); in the near area, the additional power *Add* usually increases with the age of the wearer, since the eye can accommodate less and less by itself. The optical power changes progressively from the upper far area to the near area below, allowing vision at intermediate distances. Depending on the sign of *Sph*, the general shape follows a pincushion or barrel distortion (Jalie, 2020), i.e., curving straight lines more inwards or outwards. Additionally, distortions show asymmetry between near and far areas, influenced by *Sph* and *Add*. Relating *Sph* and *Add* to perceived distortion builds a foundation for understanding PAL-induced discomfort.

Our suggested measurement is a psychophysical scale, quantify-

ing the relative change of perceived distortion for different combinations of *Sph* and *Add*. This scale indicates how much distorted a lens feels if *Sph* or *Add* changes by a certain amount of diopters. We present distortions of PALs of different near and far correction in a virtual reality (VR) simulation. Aberrations of ophthalmic lenses have been studied previously using simulations in screen-based or VR set-ups (Barbero & Portilla, 2017; Marin et al., 2008; Nießner et al., 2012; Rodríguez Celaya et al., 2005), which allow greater control over the stimulus while, in the case of VR, still allowing natural behavior. In our experiment, subjects can move freely in a virtual indoor environment, inducing distorted motion perception under natural self-motion. To study the influence of geometric distortions independent from other typical aberrations of PALs, we simulate only geometric distortions, not the blur caused by other lens aberrations.

Each trial consecutively presents three out of eleven simulated PAL distortions of various *Sph* and *Add*. Subjects responded which distortions appeared more similar (1&2 or 2&3). This ordinal data is used to fit the subject's perceptual distortion scale with ordinal embedding methods (Haghiri et al., 2020). Unlike other studies about PAL distortions, our embedding method results in an objective scaling function, which can quantify the relative influence of different lens parameters.

In contrast to screen-based distortion studies (e.g., Chandler, 2013; Charrier et al., 2007; Koenderink et al., 2017; Ponomarenko et al., 2009; Sauer et al., 2020; A. Watson, 1993), VR technology allows considering stereo and motion perception like the swim effect. It minimizes the lab-to-reality gap—increases ecological validity (Brunswik, 1955)—by recreating actual lens distortions in realistic environments in which observers can move and experience visual consequences of their own actions.

In summary, we measure the perception of geometric distortions of PALs in a realistic VR environment with natural head movements—decoupled from other typical optical aberrations. Using an ordinal comparison-based experimental paradigm in combination with an analysis by an embedding algorithm, we derive their perceptual scales individually for every observer. Our statistical modeling predicts perception across subjects well, allowing a potential application of our method for improving spectacle lenses by reducing perceived lens distortions for a generic observer.

5.2 Methods

5.2.1 Subjects

SUBJECTS WEARING spectacle lenses might already be habituated to certain distortions over the often long time having worn them. Thus only emmetropic subjects were included in the experiment to exclude this as a possible confounding factor; we assessed acuity of all subjects with the 6/6 Snellen chart. Seven male and seven female participants (mean age 24.6 years; SD 4.0 years) were confirmed not to have any known ocular diseases. One of the male subjects decided to discontinue the experiment because of VR sickness; therefore, the results of 13 subjects were analyzed. The study followed the principles of the Declaration of Helsinki and was approved by the ethical board committee of the University of Tübingen (439/2020BO). Informed consent was obtained from all participants before the measurements.

5.2.2 Stimuli

SUBJECTS LOOKED through simulated spectacles in a 3D-modelled hallway using an XTAL VR headset (VRgineers Inc, Prague, Czech Republic). The headset's horizontal FoV of 140 degree allows realistic simulation of spectacle lens distortions, affecting the part of the FoV covered by real spectacle lenses. The virtual environment was designed to replicate a scenario where distortions are visible clearly for PAL wearers: an indoor environment with many horizontal and vertical edges, shown in Figure 5.2.

Distortions of ten different PALs were included in the experiment. The far refraction ranged from -5 dpt to 5 dpt in steps of 2.5 dpt. For each of those five *Sph* values, two lenses with *Add* power 1 dpt and 3 dpt were used. All 10 lenses had the same PAL design (ZEISS Smart Life). An additional undistorted condition was included for reference. The distortions were precalculated using ray tracing based on the lens surface data provided by the manufacturer (Rojo et al., 2014). The precalculated distortions are represented as horizontal and vertical displacement of image plane coordinates; the displacement vectors are stored pixel-wise as two color channels of a texture. In the Unity game engine, the texture is used as an input to transform the rendered image by performing a coordinate transformation in a fragment shader. The procedure is performed independently for left and right eye cameras. In our experiment, the same *Sph* and *Add* corrections were used for both eyes, resulting in horizontally mirrored distortions.

Similar to the edge of a spectacle frame, we show an ellipse-shaped mask in the FoV that separates the distorted "lens area" from the undistorted periphery. The XTAL VR headset has an FoV larger than

typical spectacle lenses, which makes it possible to simulate a typical inner frame size of 54 mm by 28 mm with an ellipse of 116° inner width and a height of 80°. This frame ellipse was rendered on the image for each eye. The monocular FoV is limited in the nasal direction to 45°. Therefore, the ellipse is cut off in the nasal direction. The combined binocular perception shows a full ellipse. The ellipse thickness was 2° in visual angle.

5.2.3 *Experiment Procedure*

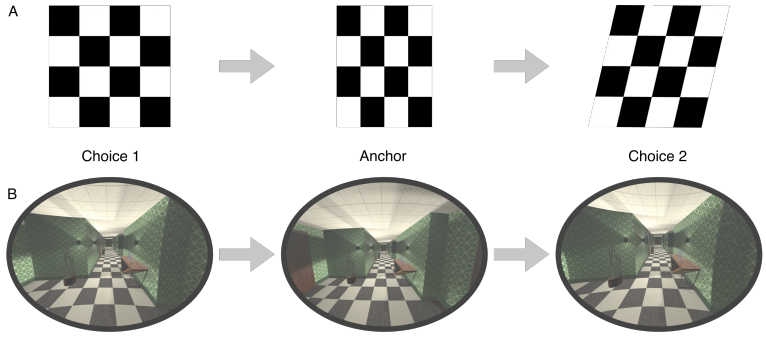


Figure 5.2: The triplet paradigm of the experiment presents three distortion stimuli consecutively in each trial: choice 1, anchor, and choice 2. The task for subjects was to answer if the first distortion (choice 1) or the last distortion (choice 2) is more similar to the second distortion (anchor). This task is equivalent to asking which of the two transitions between distortions seems smaller. A) The checkerboard pattern with simple distortion transformations demonstrates the task during the initial training phase. B) The experiment simulated PAL distortions in a virtual indoor environment. While the different distortions were presented, subjects could look around freely. Distorted motion perceived during dynamic behavior is expected to cause an unnatural or unstable perception of the environment.

The experiment used a triplet paradigm, presenting three distortion stimuli in each trial. The three distortions—choice 1, anchor, and choice 2—were all sampled from the predefined set of 11 stimuli (10 distortions and undistorted). Presentation time for each distortion stimulus was 2 s. Between stimuli, a 0.2 s transition was fading the image first to black and then to the next stimulus. Subjects had to judge the similarity of perceived distortions. They answered with a button press on a controller, which distortion—choice 1 or choice 2—appeared more similar to anchor. An alternative but equivalent instruction asks for the smaller transition: from choice 1 to the anchor or from the anchor to choice 2. In the experiment, both variants were used with the subjects.

Subjects were familiarised with this experiment procedure in multiple training phases. To clarify the experimental paradigm, in the first phase, stimuli were only checkerboard patterns distorted with clearly distinct transformations to make it easy for subjects to distinguish the distortions. One example is shown in Figure 5.2. In phase two, the stimuli showed the 3D environment of the main experiment but with trivial distortion combinations (including easy-to-perceive distortions of lenses with *Sph* 8 dpt or 5 dpt together with clearly less distorted lenses). During those two training phases, a green or red colored back-

ground at the end of each trial gave feedback to the subjects if their answer was as expected (choosing the two strongly distorted or the two clearly less distorted lenses as more similar).

In the last training phase, the distortion combinations were similar in difficulty to the main experiment. No feedback was given after the subject's answer. This training phase ensured that all subjects had learned the triplet task and were familiar with the magnitude of (real-world) lens distortions used in the subsequent experiment.

The experiment was performed in a seated position while wearing the VR headset. Subjects could move and look around freely. To induce the swim effect, subjects were encouraged to move their head. This was done by tracking head movements during training phases 2 and 3 and only continuing the trials when subjects moved their head.

THE PERCEIVED DIFFERENCES between some distortion stimuli can be small, possibly causing frustration. To increase motivation, the experiment included 20% of trivial trials, where one or two of the stimuli were exaggeratedly distorted (*Sph* 8 dpt) and the other was slightly or not distorted. Additionally, these trivial trials provide some baseline validation of subject performance because the misfitting stimulus of the three distortions is obvious and should never be chosen when following the experiment instructions correctly. In total, every subject did 413 trials, of which 83 were trivial. Each triplet combination of the 11 stimuli was presented twice, with flipped order of choice 1 and choice 2 in the second presentation. The trial order was randomized. Subjects could repeat the presentation of a trial any number of times if they felt too uncertain to respond, which might be necessary when there is a great similarity between choice 1 and choice 2 (the number of repetitions per subject can be found in Supplementary Figure B.1).

During the experiment, headset tracking data was recorded for subsequent analysis of head movement behavior. The SteamVR 2.0 tracking system (Valve Corporation, HTC; Sitole et al., 2020) was used with four base stations located around the participants' seating positions. Gaze data were captured with 120 Hz sampling frequency using the VR headset's included video-based eye tracker, accessed via the VRGineers XTAL Unity Plugin (version 2.08) and VRGineers XTAL runtime 3.0.0.77. For calibration, the manufacturer's 5-point calibration was used.

5.2.4 *Data analysis*

Psychophysical scale

THE SUBJECT'S perceived distortion was estimated from their trial responses (choice 1 or 2 is more similar to anchor) using so-called ordinal

embedding methods (Haghiri et al., 2020). These methods estimate a psychophysical scale that assigns coordinates to each stimulus so their distances agree with the subject's similarity judgments. Specifically, we used the *Soft Ordinal Embedding* (SOE; Terada & von Luxburg, 2014) algorithm, implemented in the *cblearn* Python package. The dimensionality of the scale has a great influence on how well the subjects' responses are represented in general. We chose the lowest dimensional scale that still is a good predictor of unseen triplet responses. This predictive accuracy is approximated with a cross-validation procedure (Künstle et al., 2022a). The robustness of our scale can be approximated by repeated estimates on resampled sets of responses (bootstrapping); low spread across scale samples indicates that the scale is determined well by the responses.

The perceived distortions in a scale are only a relative measure and not an absolute value——to compare scales between subjects or to create an "average observer" scale; we aligned all scales using generalized Procrustes analysis (Gower, 1975). This method minimizes the Euclidean distance between scales by iterative similarity transformations (translation, scaling, and flipping in our case) towards the mean scale. After alignment, we shifted all scales such that the origin—on average—corresponds to the undistorted lens (*Sph* 0 and *Add* 0). Accordingly, the average scaling value of the distorted lens *Sph* 5 *Add* 3 has a distance 1 to the origin.

Head movement and gaze behavior

From tracking data of headset position and gaze direction, the individual behavior during the experiment was analyzed. Since the experiment was performed in a seated position, relevant head movements are mainly rotations. We analyzed the changes in head direction by transforming the tracked headset orientation into yaw, pitch, and roll angles (Tait-Bryan angles with order y-x-z) as illustrated in Figure 5.4A. The movement velocity was computed independently for each rotation component. We calculated the mean velocity for each subject over each trial to illustrate changes in motion behavior over time. Aggregated velocity can be used to compare strategies between subjects.

Eye tracking in the VR headset is implemented independently for the left and right eye. First, the combined binocular eye gaze direction was calculated as the average of both eyes' direction vectors for each gaze sample. We compared gaze behavior between subjects by the area covered in the visual field. To do so, the binocular gaze samples were transformed to longitude and latitude coordinates in the FoV. Then, a heatmap of individual gaze distribution was calculated with a Kernel

Density Estimator of bandwidth 2 degrees as a verified upper bound of eye-tracker precision. We then calculated the solid angle of the 5% percentile, meaning the area in the heatmap, which includes 95% of the distribution's mass.

5.3 *Results*

5.3.1 *A one-dimensional scaling function models perception of PAL distortions*

ALTHOUGH physical descriptions of PAL distortions require many parameters, the perception of lenses can deviate from this parametrization. In our experiment, we use PALs of different *Sph* and *Add*, which influence the physical distortions differently: *Sph* influences more the overall shape and strength of distortions, while *Add* introduces more asymmetry in the distortion pattern. Multidimensional perception of these multidimensional distortions seems plausible. Any low-dimensional scale provides insights into which lens parameters dominate our perceived similarity of lens distortions and whether these parameters are the same in all persons. We measured individual scales of 13 subjects from triplet responses of 11 lenses of varied *Sph* and *Add*; lenses that are judged as more similar in the responses appear closer in the scale. Indeed, all subject scales are one-dimensional (Figure 5.3) and show a comparable influence of *Sph* interacting with *Add*. Additional dimensions do not increase the predictive accuracy for all subjects (see Supplementary Figure B.5). This may be regarded as surprising, given the two varied lens parameters *Sph* and *Add* and their non-linear spatial effect on distortions. Apparently, the human visual system perceives the complex PAL distortions along a single dimension only.

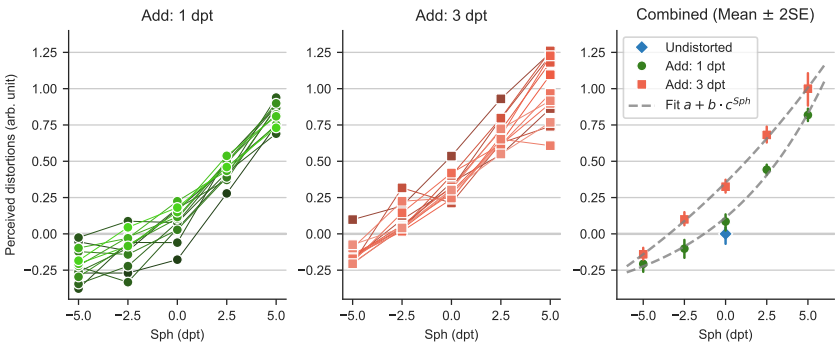


Figure 5.3: Psychophysical scaling functions of "perceived distortion" depending on the *Sph* and *Add* power of the simulated PALs. The lines in the left and middle plots show individual subjects, while the right plot shows their mean and standard error along an exponential function fit.

THE PERCEIVED DISTORTION monotonically increases with both *Sph* and *Add*. For negative *Sph* values, an increase in *Add* leads to less

perceived distortions (closer to undistorted), implying a compensation of perceived distortions; for positive *Sph*, an increase in *Add* only increases perceived distortions further.

All in all, if *Sph* is strong compared to *Add*, the shape of distortions is mainly influenced by the sign of *Sph*, leading to either pincushion or barrel distortions, which cause different perception as proven by the change of sign of perceived distortions (with undistorted at 0). The relation between perceived distortions and *Sph* can be modeled with an exponential fit of $a + b \cdot c^{\text{Sph}}$ for each value of *Add*, shown in Figure 5.3 (right), illustrating a higher rate of increase for positive *Sph* compared to negative *Sph*. Fits between *Add* 1 and *Add* 3 mainly differ in the offset a and slope b, but barely in the base c. Close to *Sph* 0, the exponential fits show a similar rate of increase for *Sph* and *Add*. With higher correction values for *Sph*, the relative influence of *Add* decreases, indicating that for high-power lenses, *Sph* dominates distortions.

The individual deviations from this exponential model do not have to be due to perceptual differences or measurement accuracy alone but can also be explained by behavior—if the distortion is perceived locally, it makes a difference where the subject looks through the lens and how they move.

5.3.2 *Head and gaze tracking reveals subjects' different behavior*

FROM REPORTS about the swim effect—an unnatural and unpleasant percept of PAL distortions during motion—we expected that especially dynamic behavior might lead to a heightened perception of distortions and thus help subjects in discriminating the stimuli. In fact, we introduced subjects to this idea by explaining the possibility of distortions becoming apparent more clearly during self-motion. Furthermore, during the training phase of the experiment, head movement was actively enforced by our experimental design. To analyze the participants' motor behavior in more detail regarding which kind of behavior they would choose to discriminate distortions, the tracking data of head and eye movements was analyzed. This allows for identifying the strategies subjects followed to distinguish distortions and test for the possible influence of behavior on the perception of distortions.

RESULTS of the head tracking show that subjects followed two different strategies: one group of 7 subjects performed continuous head movements, usually a nodding movement (pitch oscillation like in Figure 5.4B and C), some also yaw movements, while the other group of 6 subjects did not move their head or stopped after a few trials. Since subjects usually followed the same movement type throughout the ex-

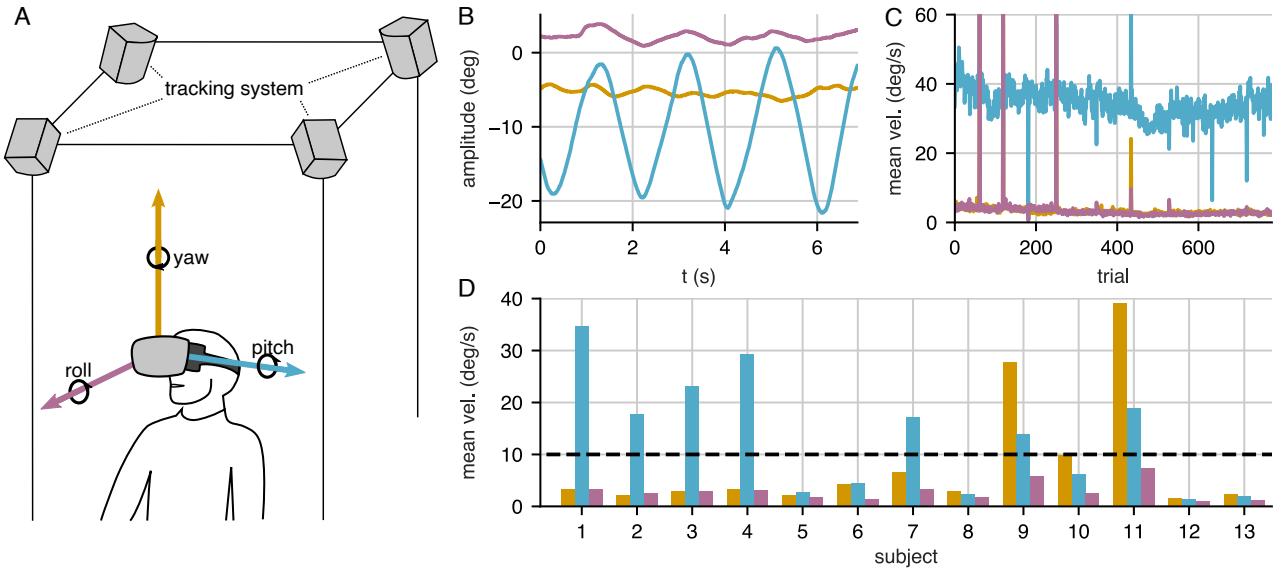


Figure 5.4: Head movements analyzed from VR headset tracking data A) Tracking setup and definition of head rotation angles. Yaw, pitch, and roll were computed from the tracked orientation of the VR headset. B) The three head rotation angles during one example trial. This example subject performed a continuous pitch movement (head nodding), while the roll and yaw angles stayed relatively stable. C) The mean rotational velocities were calculated for each head rotation component over individual trials. Data is shown for one example subject that consistently followed the same head movement behavior. D) The overview of mean angular velocities for all participants shows different head movement strategies: Some observers did not move their head, while others performed mainly a nodding (pitch) or mainly a horizontal movement (yaw).

periment (head movement over time can be found in Supplementary Figure 5.5), we calculated the mean velocity over the whole experiment as shown in Figure 5.4D.

We grouped subjects in dynamic and static observers based on their mean head movement velocity. If the mean velocity of any of the three rotation components was higher than the defined threshold of $10\,^\circ\,\mathrm{s}^{-1}$ the subject was classified as dynamic observer; otherwise as static observer. The mean roll rotation velocity never was higher than the threshold. Dynamic subjects primarily performed horizontal (yaw) or vertical (pitch) movements. The rotation velocity threshold was chosen in agreement with the examiners' observation during the experiment. The static observers had to rely only on static distortion features in the scene for their comparison judgment. During the training phase, head movements were enforced; consequently, dropping this strategy during the main experiment either indicates that the movement does not convey relevant cues for the observers or that the non-moving observers were less motivated to perform well. We compared the consistency of dynamic and static observers regarding embedding accuracy and catch-trial performance. Accuracy counts the number of triplets that agree with the estimated embedding and thus measures the general coherence of responses; responses to catch-trials, however, should be unambiguous, and any error indicates a lack of concentration. There is a significantly better embedding accuracy ($p < .05$) as well as performance in the catch trials ($p < 0.01$) for more *static* observers using the Mann-Whitney U rank test (see Figure 5.5). Consequently, it is unlikely that static observers were less motivated. Instead,

for those subjects, dynamic features contributed less to the perception of distortions. This result contrasts with the expectation that especially dynamic behavior, associated with the swim effect, would give a clear cue for distinguishing distortions and more reliable results from dynamic observers.

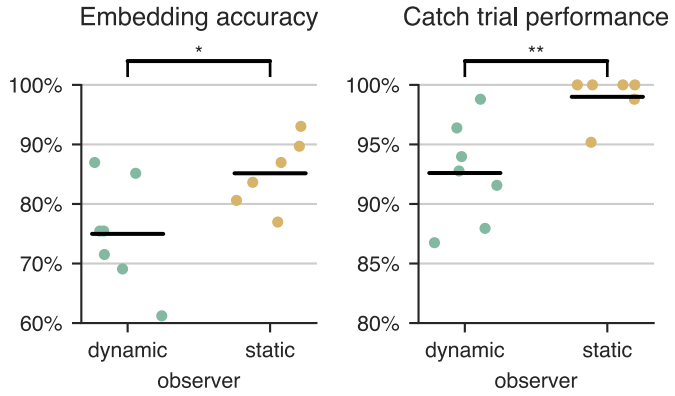


Figure 5.5: Difference in embedding accuracy and performance in the trivial catch trials between dynamic observers (head movements) and static observers (no head movements).

The distribution of gaze in the FoV for the individual subjects is shown in Figure 5.6. The gaze was mainly oriented along the center vertical axis, with variations in the latitude between individual observers. Some subjects show a high spread in gaze direction, while others stay in a more defined area of the FoV. This is reflected by the gaze area measure calculated from the individual gaze distributions. This finding suggests that some subjects, with a small gaze area, continuously fixated on the same part of PAL distortions, while others looked more at different parts of the distortion pattern. Next, we want to test if the described differences in behavior also influence the scaling of perceived distortions.

5.3.3 *Distortion perception may not be determined by overt behavior*

PAL distortions cause a complex, spatially varying transformation of the visual space, altering static and dynamic features and therefore the perception of shape, distance, and motion. Our behavior, however, influences which visual features we perceive; for example, head movements introduce perceived motion. If the various aspects of distortion perception scale differently with the PAL distortion components, then this should be revealed by differences in the scaling function between subjects of different behavior. Especially the difference between static and dynamic observers should show how the swim effect, associated with dynamic situations, contributed to distortion perception. But also, different gaze strategies might cause differences in perceived distortion and thus in the recovered perceptual scales. Eye tracking re-

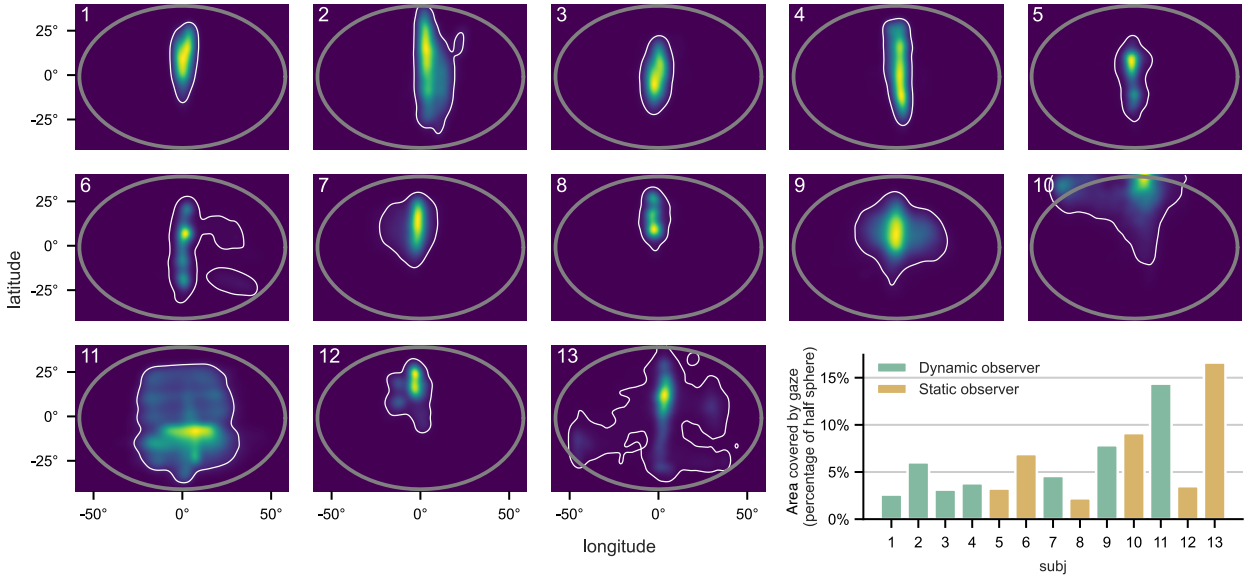


Figure 5.6: Individual gaze distribution of all observers with Gaussian kernel smoothing. Binocular gaze samples (in head-relative coordinates) of the whole experiment were combined to calculate the gaze distribution in the observers' FoV. The grey ellipse represents the virtual frame size, which was visible as a black border during the experiment. The white contour encompasses 95% of the gaze distribution mass. The area enclosed by this contour for individual subjects is shown in the bar diagram as percentage of the half-sphere area.

sults revealed differences in the spread of gaze. A wider area of gaze implies that subjects see a higher variability in distortions.

To statistically evaluate this possible influence of behavior on the scaling function, we used a linear model to predict perceived distortions depending on *Sph*, *Add*, and head and gaze behavior. For the two head-movement groups, we introduced a categorical variable in the model; gaze spread is modeled by the area in the field of view, covered by gaze (according to 95% KDE distribution mass, see Figure 5.6). The model's fixed effects include *Sph*, *Add*, head-movement group, gaze area, and all their interactions. The preceding Procrustes analysis aligned the subjects' individual scales already by shifting and scaling so that random effects for slope or intercept do not have to be considered in the linear model.

To fit the linear model, we used `stattoolbox`'s `ols` function in Python. *Sph* and *Add* both show significant effects on perceived distortions ($p < 0.001$). No other effects or interactions are significant. The differences in head movement and gaze behavior did not lead to significantly different perceptions of distortions, which indicates that perceptual effects of both static and dynamic features scale similarly with the amount of PAL distortions.

This result is, from a practical point of view, very good news: It implies that the general scaling of distortion perception is similar for all observers, independent of their specific and often idiosyncratic head

and eye movements; different behavior does not lead to differently perceived distortions, which in turn allows a general quantification of perceived distortions for a specific PAL.

5.3.4 *A non-linear fit of scales predicts perception of PAL distortions*

ONLY IF we can make predictions about the distortion perception of unmeasured subjects—based on scales measured from other subjects—is there a possibility that perception models can actually be used to improve lens designs beyond individual designs. We assessed the predictive ability of three differently complex regression models in a leave-one-subject-out procedure: the models are trained on all but one subject's scaling functions and subsequently tested on the omitted subject. We found that the perception of most subjects follows the same regularities, well captured by an exponential model.

THE ASSESSED MODELS include the exponential fit from Figure 5.3 along with baseline and ceiling performance models to provide a reference. The baseline model is a linear regression, and the ceiling model is a random forest regressor (Breiman, 2001), known for excellent out-of-the-box performance in non-linear problems. R^2 scores in Figure 5.7 show an overall high predictability in all but the baseline model, indicating that most of the subjects' scales can be predicted accurately.

Inspecting the scales with lowest R^2 scores (compare Supplementary Figure B.4), we see some individual differences. In the extremes, subject 1 shows a noticeably higher influence of *Add* and a more pronounced flattening in negative *Sph*, while the *Add* parameter has almost no influence in the scale of subject 13. The responses of subjects 4 and 11 during the experiment seem less consistent, resulting in higher variations of the scaling estimates. For these subjects, the collection of additional trials might improve the agreement with other subjects and, therefore, improve predictability.

FOR MOST SUBJECTS, the exponential and random forest models substantially increase prediction over the linear model, again underlining the non-linear influence of *Sph* and *Add*. The random forest, however, can use its greater flexibility with only one subject to predict the scale better——an exponential model actually seems to describe the relationship between perceived distortions and lens parameters very well.

5.4 *General Discussion*

WE PERFORMED a VR experiment to determine a psychophysical scale of perceived optical distortions of PALs. Distortions were simulated in

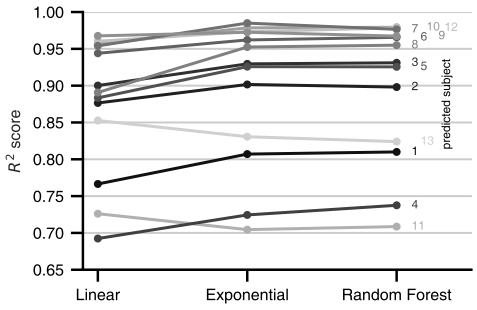


Figure 5.7: Performance in predicting a subject's perceived distortion with models that are trained on the remaining subjects' scales along *Sph* and *Add*. The linear and random forest models indicate baseline and ceiling performance. Exponential functions of *Sph*, grouped by same *Add*, predict similarly well as the much more general random forest model.

VR based on precalculated ray tracing of real PAL designs. A triplet paradigm was used to retrieve ordinal data of the relative perceived distance of distortions. The approach was tested with a set of PALs of varying *Sph* power (far correction) and *Add* power (near correction). In our explorative study, observers could move their heads and eyes freely to induce the perception of unnatural motion (swim effect) associated with the distortions of progressive lenses.

The scaling functions retrieved by fitting the ordinal data show a similar trend for all observers: perceived distortions increase with spherical power. For positive *Sph*, increasing the *Add* power results in stronger distortions. In contrast, the positive *Add* power in combination with negative *Sph* power reduces the perceived distortions (closer to undistorted), suggesting an advantage for short-sighted PAL wearers. For very high or very low *Sph*, the relative influence of *Add* on perceived distortions seems less relevant; close to *Sph* 0, *Add* and *Sph* seem to have a similar influence on perceived distortions. The relationship is very well modeled by a simple exponential function.

Behavior measurements of head and eye movements show that observers followed different strategies in the experiment. The most apparent strategy difference is the head movement behavior, which also leads to differences in the performance in catch trials. Some subjects moved their head continuously (nodding or shaking motion) while others kept their head stable, despite being instructed in the training of the experiment to perform head movements. This indicates that some observers might be less sensitive to perceive distortions from optic flow. This also agrees with the experience of PAL wearers: some individuals are more sensitive to the swim effect than others (Alvarez et al., 2009). A speculative explanation would be that subjects who relied on head movements perceive the influence of distortions on optic flow more clearly and might suffer more from the swim effect.

However, no influence was found by head movement or gaze be-

havior for the psychophysical scaling function. We conclude that the global distortion pattern, not only the local distortions, influences the perception of distortions with static and dynamic features. For our virtual scene, distortions perceived from static or dynamic features scale similarly to the PAL parameters. Additionally, the scales generalize well to new observers, confirming that our psychophysical scale can be used as an objective quantification method for PAL distortions. The remaining differences between subjects' scaling functions (for example, differences in the relative influence of *Add* power) could result from individual differences in perception or behavior. We suggest follow-up experiments focusing on increased stimulus and behavioral control. One aspect of our experiment that would be worth investigating in a more controlled setting is the significance of binocular vision for the swim effect. VR technology offers the opportunity to control binocular vision, which helps in investigating the influence of distortions on depth and surface curvature perception. It might be vital to perform binocular tests to ensure suitable participant recruitment. In any of those potential follow-up experiments, an increased number of participants would provide valuable information to identify potential patterns in the perception of observer groups, which could be included in our perceived distortion model.

Further research should investigate static versus dynamic effects in detail as well as extend our inquiry to other lens aberrations. Our study concentrates on the influence of distortions on perception. For real PALs, spherical, astigmatic, and higher-order aberrations negatively impact vision. The possible influence of those aberrations and the interaction with distortion perception could be studied in future experiments by including realistic blur in our VR simulation (Sauer, Wahl, & Habtegiorgis, 2022). In more realistic viewing conditions, PAL wearers will adapt their behavior to gaze through the clear areas of the lens (Rifai & Wahl, 2016). This development of a "head-mover" behavior (Hutchings et al., 2007) should not be confused with the dynamic movement strategy of one part of our subjects. In our experiment, the subjects' behavior reflects their strategy to distinguish differences in the distortions, while in everyday life, PAL wearers would likely follow an approach that minimizes discomfort. The gaze distribution would be determined by the current activity. For instance, during walking, the gaze would be directed downwards. In that sense, our results for behavior do not reflect the influence of PALs in everyday life but individual strategies for distinguishing between distortions.

5.4.1 Conclusion

In this study, we introduced a new measurement method for determining a psychophysical scale of optical distortions. Our measurements show a high agreement between subjects, allowing predictions of PAL distortion perception in general. These results reveal the potential of using psychophysical methods for understanding the swim effect and could help to improve future optical designs of PALs. As stated by the Minkwitz theorem, a total reduction of aberrations in the lens is impossible. Design choices for a given correction power can only change the spatial distribution of aberrations. Choosing a design with a lower amount of perceived distortions can contribute to reducing distortion-related discomfort for PAL wearers and increase satisfaction. To quantify different designs for their perceived distortion, it is required to repeat our experiment to measure perception not only depending on the correction power (*Sph* and *Add*) but directly on parameters describing the possible differences in PAL designs for a given correction. With a model based on the results of this suggested experiment, an arbitrary PAL design could be quantified for perceived distortions purely based on ray-traced distortion data without testing it in an additional experiment. As a completely new approach to lens design, this perception-focused optimization might lead to a realignment of current lens design processes.

Code and data availability

The dataset generated during this study as well as the analysis code is available on GitHub[2].

[2] https://github.com/ZeissVisionScienceLab/PAL-distortion-scaling

6 Discussion

We generalize from one situation to another not because we cannot tell the difference between the two situations but because we judge that they are likely to belong to a set of situations having the same consequence.
— Roger N. Shepard
(Shepard, 1987)

In this thesis I have introduced a machine learning pipeline for psychophysical scaling with ordinal comparisons. This pipeline consists of a software toolbox (chapter 2), a procedure for determining the dimensionality of the scale (chapter 3), and an analysis of its quality and stability (chapter 4). Finally, I presented an application of the pipeline in a study to measure the lens distortion of glasses simulated in virtual reality (chapter 5).

In this chapter, I first discuss the strengths and weaknesses of the pipeline: The role of machine learning methods for scientific discovery, the challenges of methodological recipes, and the coverage of different pipelines for psychophysical scaling. I conclude the chapter with suggestions for future research on additional approaches to comparison-based machine learning in psychophysics and potential applications of our pipeline in academia and industry.

6.1 Machine learning as a key ingredient

In the previous chapters I present the components of a pipeline for estimating psychological scales from similarity comparisons using machine learning algorithms. This raises the question of whether it is really necessary to use machine learning in the established field of psychophysics. In chapter 1 I listed specific benefits of ordinal embedding algorithms over traditional scaling methods: They are theoretically well-studied, flexible in how triplets are selected, robust to noise, and can create multidimensional representations. Here, however, I discuss the role of machine learning as a methodological and technological tool.

Psychophysics is inherently interdisciplinary and always relies on cooperation between psychology and methodological fields. Multidimensional scaling (MDS) was, from the beginning, a team effort of mathematicians, cognitive scientists, and psychologists (Kruskal, 1964a, 1964b; Shepard, 1962; Torgerson, 1952) with substantial theoretical contributions from statisticians (Hefner, 1958; Ramsay, 1969;

Suppes & Zinnes, 1963). Many of the application questions addressed in this thesis have already been considered in the context of MDS. For example, there is also the question of the correct embedding dimensionality (Borg & Groenen, 2005; Davison & Sireci, 2000; Oh & Raftery, 2001; Spence & Graef, 1974), as well as the estimation of variability (Gronau & Lee, 2020; Jacoby & Armstrong II, 2014; Kiers & Groenen, 2006; Zinnes & MacKay, 1983).

Machine learning is just the latest tool in the scientific toolbox for analyzing data and building on mathematics and statistics. This tool comes in the form of stochastic algorithms that learn patterns from data without being explicitly programmed (Samuel, 1959). This makes them particularly useful for the analysis of behavioral data, which inherently contains a degree of randomness.

In psychophysical scaling, the distribution of the inner space is often unknown, which is why, contrary to classical statistical models, learning from data with ordinal embedding methods leaves the necessary freedom for explorations. We took an explorative perspective in our study on the perceived distortion of varifocals in chapter 5, where little prior knowledge was available, and our embeddings surprisingly predicted most of the subject's responses in one dimension.

Machine learning methods' flexibility is also a huge challenge for scientific knowledge gain. Science is about more than the fit to data, but about the evidence for (or against) theories. On the one hand, this requires us to determine the quality of the model (cf. chapter 4) and then derive insights. For this reason, in chapter 3, we have presented a procedure which can determine the dimensionality of the scale and provides a statistical interpretation aid. The procedure indicates the minimum dimensionality at which the scale is likely to predict observers' responses best—we suggest interpreting this dimensionality as a lower bound on the dimensionality of the inner space, closing the loop back from a statistical method to scientific knowledge gain.

Besides the algorithmic properties, one of the greatest technical strengths of machine learning is the open-source software infrastructure. Our `cblearn` toolbox, presented in chapter 2, is a machine learning package for working with ordinal comparison data that can be applied to psychophysical scaling. This machine-learning perspective ensures the library appeals to a broader range of users, including both psychophysicists and computer scientists. While psychophysicists can actively use the procedures in `cblearn` to analyze their data (Huber et al., 2024; Sauer et al., 2024), computer scientists can benefit from easy access to real datasets and fast algorithm implementations to benchmark new methods (Mandal et al., 2023). A broad user range is essential for long-term maintenance as an active open-source project and helps to

integrate sustainably into the scientific and machine-learning ecosystem.

This open-source ecosystem is a major driver for advanced model implementations. I could not have implemented this toolbox as extensively and efficiently as a one-person project without relying heavily on high-quality open-source projects like `numpy`, `scipy`, `scikit-learn`, or `pytorch` (Ansel et al., 2024; Buitinck et al., 2013; Harris et al., 2020; Virtanen et al., 2020) with fast matrix processing or automatic differentiation. This foundation, as well as the machine learning-oriented API, allows seamless collaboration with other packages. All chapters of this thesis make use of the ecosystem integration: We used pre-implemented cross-validation and grid search methods for the dimension estimation in chapter 3, metric and sampling methods for the quality estimation in chapter 4, and all of them combined in the data collection and analysis of the distortion scale experiments in chapter 5.

However, communication and specialized vocabulary remain a major challenge in our interdisciplinary work between machine learning and psychophysics. In the publications, we translate terms appropriately to the terminology from the journal's field; this is not possible in our `cblearn` package that is used by both communities, such that the API uses technical terms that do not necessarily match the psychological concepts. For example, in the library *scales* are called *embeddings* and *trials* are called *queries*. In addition, the API itself might feel more technical as we chose a modest degree of abstraction. While, for example, MLDS (Knoblauch & Maloney, 2008) provides a specific function to determine the scale's standard deviation with bootstrapping, our users would have to build this functionality themselves in three simple steps (resampling, alignment, and standard deviation). The level of abstraction always imposes a challenge for a broad user range. Despite aiming for a good trade-off in our solution, there is potential for a high-level API specifically designed for psychophysicists. Such a potential specialized interface ought not to replace `cblearn` but build on its foundation, providing frequently used functions to lower the learning curve.

6.2 *From algorithms to recipes*

JUST AS a piece of wood and a blueprint are not enough to build a table, no one can gain knowledge about perception with experimental data and an embedding algorithm only. The carpenter needs a saw that he can adjust precisely and a measuring tape to evaluate the quality of his work; we need algorithm implementations that are easy to use, instructions on the algorithm configuration, and methods to assess the quality of the resulting scale (cf. Haghiri et al., 2020).

With the `cblearn` toolbox, introduced in chapter 2, I intend to equip psychophysicists with the tools to estimate ordinal embeddings but to load, simulate, and preprocess comparison data and to evaluate the resulting scales. New users will find detailed documentation and examples for their first steps. I have presented a recipe for configuring dimensionality in the form of a statistical procedure in chapter 3, and recipes for evaluating the quality and variability of the scale in chapter 4.

The recipes always come with detailed instructions and extensive applications in simulations and behavioral data. Their description is much more verbose than the "rules of thumb" in Haghiri et al. (2020), and the validations much more extensive than for the recipes in Knoblauch and Maloney (2008, 2012a). This raises the question of whether such elaborate recipes are necessary or whether I am overdescribing trivial things here. It remains difficult to quantify whether a community needs such detailed instructions, and yet there are certain indications of their usefulness. In psychometrics, for example, the works of Wichmann and Hill (2001a, 2001b) or Schütt et al. (2016) are highly influential on the community and among the most-cited publications of these authors (according to scholar.google.com in July 2024). These papers show how psychometric functions should be fitted and their variability determined, and provide the necessary software solutions. The topics are similar to those discussed in this thesis, but target psychometric functions instead of comparison-based psychophysical scales.

Recipes for the use of (new) methods seem to give confidence in their application but also come with the risk that users rely too much on them and overlook missing assumptions or over-interpret results. In psychology, this is a well-known problem when applying statistical tests. For example, naïve users take effects as confirmed just because the p-value is less than 0.05, even if distributional assumptions are violated, or the effect size is too small to impact the real world (Benjamin et al., 2018; Wagenmakers, 2007)—and if the p-value is too big, the instructions could be varied to obtain the "right" result (Head et al., 2015; Stefan & Schönbrodt, 2023).

This danger persists with the presented recipes for estimating dimensionality or variation in that both can be misinterpreted as absolute values. As shown in the simulation experiments in chapter 3, the estimated dimensionality is a lower bound and can vary depending on the data, the embedding algorithm, and the cross-validation settings; similarly, the variability depends on the noise function and the prior. As in any modeling solution, a balance must be struck between

oversimplification in the instructions and differentiation in the results.

6.3 *Constructing a scaling pipeline*

EVEN the best model is worthless if it is trained with data from a poorly controlled experiment, so neither data collection nor analysis should be approached separately. A scaling pipeline provides holistic methods for the entire process. Specifically, the scaling process begins with experimental design and the collection of ordinal comparisons. This is followed by dimensionality determination and embedding before the scale is further validated, visualized, and interpreted to draw scientific conclusions.

WE DEMONSTRATED the entire process in a practical application with the psychophysical study in chapter 5. Since embedding algorithms have already been discussed in other works (Agarwal et al., 2007; Haghiri et al., 2020; Tamuz et al., 2011; Terada & von Luxburg, 2014; van der Maaten & Weinberger, 2012), we have focused on the practical steps of the analysis process in chapter 3 and chapter 4. During the whole pipeline, our toolbox `cblearn` (chapter 2) can be used. It provides utilities to sample random triplets to be collected in the experiment and to load the resulting response files. In addition, implementations of embedding algorithms and our dimensionality estimation procedure simplify the scale estimation. This scale can be further visualized and analyzed (e.g., cross-validated) with interfaces to the scientific Python ecosystem.

MULTIPLE scientific groups developed methods for psychophysical scaling with ordinal comparisons, but they focus on different aspects as visualized in Figure 6.1. "Pure" embedding methods such as SOE (Terada & von Luxburg, 2014), GNMDS (Agarwal et al., 2007), or t-STE (van der Maaten & Weinberger, 2012) are only a small part of the overall study. The analysis of the scales can be improved by integrating the dimensionality estimation directly into the embedding algorithm (e.g., `psiz`; Roads & Love, 2021; Roads & Mozer, 2019), or by fixing it to 1D as in MLDS (Maloney & Yang, 2003). MLDS is additionally supplemented with validation methods and variability estimation via bootstrapping (Knoblauch & Maloney, 2008, 2012a). Haghiri et al. (2020) already shared ideas on how to build on the pure embedding algorithms for multi-dimensional scales to estimate dimensionality, quality, or variability and influenced large parts of this work's pipeline with their ideas.

Other works focus more on collecting the data. Packages such as `salmon` (S. Sievert, 2021) and `NEXT` (Jamieson et al., 2015) that pro-

vide methods to select the most informative trials actively are valuable when collecting large datasets (based on, e.g., CKL; Tamuz et al., 2011). An even more comprehensive helper for data collection is the Meadows platform (Meadows Research Ltd), which provides prebuilt scaling experiments in a no-code environment. Although Meadows offers triplet experiments, it cannot automatically generate scales for them.

As the pipeline of this work does not yet cover much of the data collection part, I see great potential in creating interfaces. For example, `salmon` can actively collect triplets, and in a second step, `cblearn` can determine dimensionality and estimate scale. If these methods could be integrated into no-code platforms such as Meadows, a new audience could benefit from the advantages of comparison-based experiments. By providing such a connection through software interfaces in the framework, individual projects could play to their strengths and allow researchers to use the methodology as efficiently as possible from start to finish.

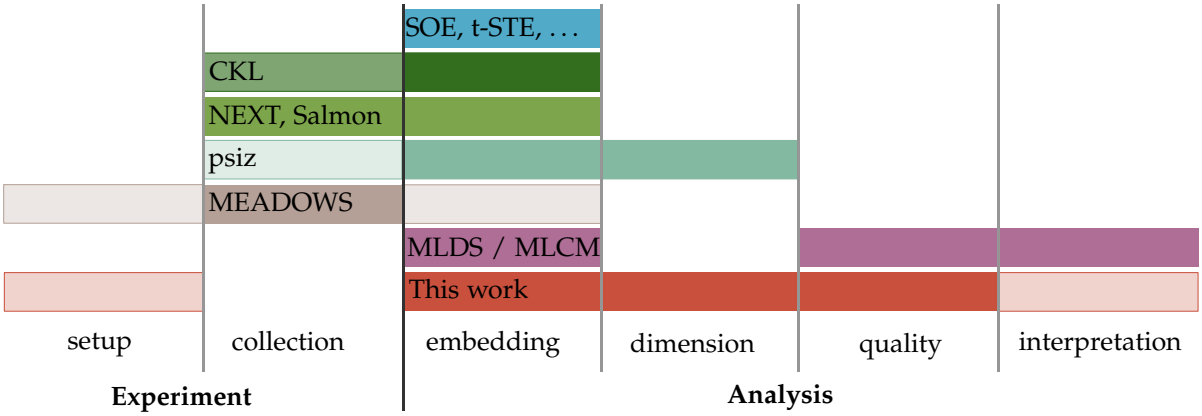


Figure 6.1: Algorithms and pipelines for psychophysical scaling from ordinal comparisons. The projects prioritize different phases for which they provide methods (colored) or only hint at them (shaded). References to the projects can be found in the main text.

6.4 *Future research directions*

The previous chapters presented a fundamental pipeline for psychophysical scaling with ordinal comparisons for the majority of studies[1], using ordinal embedding methods. However, there are other machine learning algorithms that are promising candidates to solve further challenges of comparison-based scaling. This section structures those promising methods by three key questions: How can scales be collected more efficiently? How can perceived similarities be represented if Euclidean distance appears inappropriate? And how can scientific insights be gained from similarity representations? Here I outline possible directions that we can also find as methods in the comparison-based machine learning literature beyond ordinal embedding.

[1] In line with the Pareto principle (Pareto, 1896).

6.4.1 Trial selection procedures

In general, a scaling experiment attempts to measure the perception of as many stimuli as possible. However, when collecting random triplets, the required number of trials grows super-linearly[2], which severely limits the number of stimuli that can be collected in practice. Therefore, here I discuss different approaches for more trial-efficient scaling approaches.

[2] Haghiri et al. (2020) recommend at least $DN \log N$ trials for D dimensions and N stimuli.

A common solution to reduce the number of trials is presenting them not randomly but prioritized by their "information content". These methods are known as *adaptive sampling* in psychophysics (e.g., the Quest sampler for threshold methods; A. B. Watson & Pelli, 1983) and in machine learning as *active learning* (Settles, 2009). Several papers describe active learning methods for ordinal embedding algorithms, intending to accelerate crowd-sourcing studies (Jamieson et al., 2015; Roads & Mozer, 2019; S. Sievert, 2021; Sievert et al., 2023; Tamuz et al., 2011). Although these methods have already been used successfully in extensive online studies (Lagunas et al., 2019; Roads & Love, 2021), their practical added value remains controversial: In simulation studies, the required number of triplets does not decrease (Jamieson et al., 2015), or only decrease after an initialization period (S. Sievert, 2021; Tamuz et al., 2011). Roads and Mozer (2019) report that active sampling saves 57% of the experimental time compared to randomly selected trials. However, even this near doubling of feasible trials comes at the cost of making an adaptive experiment technically challenging because the selection mechanism has to run during the experiment (S. Sievert, 2021, ch. 4), and the selection criteria may unintentionally bias the result (cf. biases in adaptive psychometric function fitting; Beck & Shaw, 1965).

Laboratory studies can collect only a fraction of the trials that can be collected in online studies, making it even more difficult to initialize a model for active trial selection from collected data. Instead of data, however, the experimenter's experience, pilot studies, or literature can also provide certain prior assumptions about a scale. The scaling algorithm of Heim et al. (2015) is a promising candidate to speed up trial collection by exploiting those prior assumptions as so-called *auxiliary information*. The auxiliary information guides trial selection and should speed up data collection as long as the information shows roughly similar trends to the real scale.

Looking through the eyes of an experimenter, one misses the role of the observer in active learning methods. Although selecting the most informative trials may be *mathematically* optimal, it might not be *ex-*

perimentally optimal. Human subjects can get tired or frustrated such that skilled experimenters assure that the difficulty of the trials is balanced. In adaptive sampling, however, we must assume that the most informative trials are all non-trivial to answer. Additional biases from serial dependencies are known in adaptive psychometric function fitting (Kaernbach, 2001) but understudied in active sampling of ordinal comparisons. Future work might focus on studies investigating potential biases but also on new sampling strategies that consider the task difficulty.

THE EXPERIMENT DURATION is influenced not only by the stimuli shown but also by the question asked. Experimental tasks usually combine several comparisons in a single question (please find an overview on ordinal comparison tasks in chapter 1). For example, an 8-rank-2 task is equivalent to 13 triplet questions by presenting a 3-by-3 grid of stimuli and asking subjects to rank two stimuli most similar to the centered. Roads and Mozer (2019) demonstrate a 8-rank-2 experiment that is about 3× faster than a corresponding triplet experiment. However, it is well-known in experimental psychology that the instructions can bias the results (e.g. Robinson, 1976). In a Bachelor's thesis I have supervised we have observed evidence that observers use different judgment criteria for odd-one-out comparisons than for triplets, even though both tasks can present identical stimuli and can mathematically result in identical comparisons (Schönmann, 2021; Schönmann et al., 2022). Future studies are required to disentangle the benefits of more complex comparison tasks on experiment duration from their potential biases on the resulting scale.

6.4.2 Non-Euclidean similarity measures

THE SCALES in this work are based on the standard Euclidean distance to measure the dissimilarity in the internal space, as does most of the ordinal embedding literature. While Euclidean representations are easy to visualize and post-process with, for example, clustering algorithms, depending on the stimulus, it may be necessary to use other metrics or even non-metric similarity measures to represent observers' responses adequately. Here, I would like to present machine learning solutions to generate non-Euclidean representations from ordinal comparisons, which, in my opinion, should be applied in psychophysics in the future.

Several psychological studies have found that scales fit their data better when they do not use an Euclidean metric; however, the "better" metric varies from stimulus to stimulus (Attneave, 1950; Shepard, 1964; Waraich & Victor, 2024; Zhou et al., 2018). Logvinenko

and Maloney (2006), for example, find that their similarity measures of lightness perception are represented better with a *city block* instead of Euclidean distance. Implementing ordinal embedding algorithms with alternative distance metrics is straightforward using automatic differentiation techniques like the PyTorch implementations in *cblearn* (chapter 2). The metric can be selected with hyperparameter optimization procedures such as the grid search, which we use in the dimensionality selection procedure in chapter 3.

In some experiments, however, the observers' response criteria might vary between trials, for example because object images are compared based on the trial's dominant feature, such as size, number, or color. As the weighting of these features may vary on a trial-by-trial basis, no metric representation satisfies the comparisons. In these cases, so-called multiview triplet embeddings (MVTE; Amid & Ukkonen, 2015) might be a valuable solution by simultaneously learning multiple scales, one for each property. Each triplet contributes to one or various scales with different weighting.

For researchers primarily interested in the distance metric rather than the scale, machine learning algorithms can determine this metric directly from ordinal comparisons. In that case, it can typically be estimated in the form of a Malahanobis distance matrix from the triplets or quadruplets using metric learning algorithms (Liu et al., 2012; Shi et al., 2014; Weinberger et al., 2005). Perrot et al. (2014) used determined local distance measures between colors in the RGB image space of cameras. Metric learning on behavioral data could be valuable in investigating open questions about color perception, for example the unique role of yellow and brown[3] (Bartleson, 1976)

[3] Most hues are perceived as such regardless of illumination—light blue and dark blue are both perceived as blue—which is why some color models have an independent lightness dimension (find an introduction to color models in Fairchild, 2013). However, these models are poor at describing yellow, as dark yellow becomes brown, and light brown becomes yellow, a unique outlier in our understanding of color perception (Vincent, 2017).

In the extreme perceived similarity of some stimuli is described poorly by any metric (cf. review about models of perceptual spaces by Roads & Love, 2024). The corresponding similarities do not fulfill the metric axioms such as symmetry or triangle inequality (Tversky & Gati, 1982). We find such non-metric perceptual spaces, for example, in animals or things that can easily be arranged hierarchically. When each animal corresponds to a leaf in a tree, perceived similarity can be represented by path length (Sattath & Tversky, 1977). These representation trees can—in theory—be constructed from triplets (Victor et al., 2023). Methods to construct such a tree from data are so-called hierarchical clustering algorithms, such as the triplet-based ComparisonHC algorithm (Ghoshdastidar et al., 2019).

6.4.3 Predictive and functional models of behavior

MOST OFTEN, we estimate psychophysical scales not just to measure a subjective intensity but to learn about their relation to stimulus properties (Fechner, 1860; Gescheider, 1988; Stevens, 1957). Scales such as those presented in this thesis, however, measure stimuli in isolation without establishing a direct relationship to the triggering intensity and without modeling the intensities between stimuli. For this reason, we will discuss here how to construct real, continuous scaling functions from these scales.

REGRESSION MODELS can learn the relation between stimulus parameters and perceived intensities, as demonstrated in chapter 5. They can be used to interpolate perceived intensities for stimulus parameters not observed in the experiment. Probabilistic regressors, such as the Gaussian processes in chapter 4, can provide information about the uncertainty of these predictions.

Future research might use standard methods from explainable or interpretable machine learning (XAI/IML) to reveal the functional relationship between stimulus properties and perceived intensities in these (non-linear) regression models (Freiesleben et al., 2022). For example, feature importance analyses (e.g., Breiman, 2001; Hooker, 2007; Sobol, 1993) can express which parameters influence perception, and partial dependency plots (Friedman, 2001) can visualize the relationship of parameter values. In chapter 5, we introduced regression models that predict perceived distortion from lens properties. Follow-up work on this research can profit from those XAI methods to investigate the importance of physical lens properties on perceived distortion to understand which lens characteristics affect the wearer the most.

However, further work on psychophysical scaling might combine ordinal embedding and regression to a single model. One could train a deep learning regressor with a triplet-based loss to learn the ordinal embedding by incorporating the stimulus properties. We already know similar learning methods as contrastive learning of image classifiers (Chen, Kornblith, Norouzi, & Hinton, 2020).

EVEN MORE insight into the stimulus-perception relationship could be gained by building not only statistical but functional models that explicitly combine factors, e.g., additively and multiplicatively. Conjoint measurement models like MLCM explicitly combine additive and interactive effects of perceived stimulus dimensions (Ho et al., 2008) to predict behavior. Statistical comparisons of (nested) model alternatives provide additional insight. These comparisons help to understand which parameters contribute to behavior and if they interact.

Unfortunately, MLCM cannot be trained on arbitrary triplets, but future work could extend ordinal embedding towards conjoint measurement as described in the following.

An ordinal embedding-based conjoint measurement algorithm could be created with few changes to standard ordinal embedding methods. First, we assume that the dimensionality of the embedding is fixed by the parameters to be analyzed. Instead of representing each stimulus by separate coordinates, they share coordinates based on the corresponding stimulus parameters: Stimuli with identical parameter values share the corresponding coordinate dimension. This coordinate sharing is straightforward and can be implemented as weight-sharing. In the objective function, the Euclidean distance is replaced by the functional composition of parameters. For example, an additive model could be implemented by the city-block distance.

6.5 Potential applications

Rather than exploring further methods, the tools presented in this thesis already offer potential for a wide range of applications in research and development and everyday life.

Many psychophysical studies already profit from the advantages of comparison-based experiments for 1D scales with MLDS (find examples in chapter 1). Still, multidimensional scales, as discussed in this thesis, are required to answer research questions today. State-of-the-art technology enables psychophysical experiments that more closely mimic real-world settings, making findings relevant to our everyday lives (cf. *ecological validity*; Brunswik, 1955). For example, our virtual reality setup described in chapter 5 enabled the measurement of dynamic distortions changing with motion as with real glasses, rather than static distortions as in previous screen-based studies (e.g., Sauer et al., 2020). However, the increased realism also means that the effects on perception can be significantly more complex, and multidimensional modeling may be necessary (but do not have to; see the discussion in section 3.4).

The fact that the triplet method can even be applied to dynamic, sequentially presented stimuli, as I have shown in chapter 5, further increases the application possibilities for investigating motion perception, auditory perception, depth perception (motion parallax), or distortions in images and videos. The quantification of perceived video and image quality would benefit significantly from applying the methodology presented in this thesis. Instead of evaluating the quality of compression methods by rating procedures (so-called Mean Opinion Scores, e.g., see Ponomarenko et al., 2009), a comparison-based scale

could provide more objective estimates (Chandler, 2013; Charrier et al., 2007; Men et al., 2021, cf.).

Beyond pure research, many industrial applications can benefit from the precise, multidimensional measurement of visual perception. As described in chapter 5, subjective optical properties of glasses, binoculars, and microscopes could be optimized during development using a perceptual metric based on a psychophysical scale. Similarly, photography and movie postprocessing, or material and illumination design, all rely on accurate, multidimensional models of human perception——and most often, it's a challenge to obtain them objectively (cf. Pousset et al., 2010; Zamir et al., 2021). Usually, elaborate heuristics are created (cf. Burley & Studios, 2012) instead of measuring the stimuli psychophysically. We have shown that this measurement is possible in a master's thesis that I supervised: The user interface of photo editing software provides multiple sliders to manipulate qualities related to image contrast. Hölscher (2022) measured the perceived intensity of those sliders and analyzed whether they were perceptually equidistant and orthogonal or if it would be possible to construct a more intuitive set of sliders.

In addition to visual perception, ordinal embeddings can also be used to measure the perception of linguistic concepts and could become a valuable quantitative methodology in fields like philosophical and social science. As a case study in experimental philosophy, we measured the similarity of concepts in the context of "truth" in an online triplet experiment and compared them with philosophical *truth* theories (Huber et al., 2024). Even three months after the experiment, we were able to predict $\approx$ 70% of individual subjects' responses to a text exercise on truth comprehension using their scales.

Everyday applications of scaling methods can be found in peer evaluation problems. In large online courses, so-called MOOCs, peers are asked to evaluate their fellow students' exercises as objectively as possible (Caragiannis, 2017; Sajjadi et al., 2016). This could be done more intuitively using an ordinal triplet comparison between three submissions, and the grading scale could be computed using a 1D embedding. Similarly, the method needs to revolutionize the field of product and service ratings. Most websites currently ask users to rate t-shirts, hotels, or restaurants with one to five stars. This rating is far from an objective, equidistant quality evaluation but suffers from the biases of direct scaling methods (see chapter 1). However, ordinal comparisons of products or service triplets can easily be pooled across customers to obtain an objective 1D scale.

6.6 *Concluding remarks*

THIS THESIS provides a pipeline for applying machine learning methods for ordinal comparisons to psychophysical scaling practice, enabling an intuitive and flexible, yet objective, measurement of a person's inner space.

I provide a software toolbox that allows psychophysicists without knowledge of machine learning to use comparison-based algorithms to infer psychophysical scales. The toolbox is the foundation for comparison-based procedures beyond pure inference. These include the procedures presented in previous chapters that help researchers interpret scales by estimating dimensionality and quantifying quality and stability. Interpretation transforms data and methodology into knowledge. This relationship is demonstrated in the final chapter with a novel application of scaling methods to the perceived distortion of varifocals.

In summary, we equip psychophysicists with the tools to successfully apply ordinal embedding algorithms as robust multidimensional scaling methods. Thus, this work ensures that machine learning algorithms with ordinal comparisons will make relevant contributions to new insights in perceptual research and beyond.

A Supplementary dimensionality procedure material

The content of this chapter equals the supplementary of our published article[1]. The author's contributions are described in chapter 3.

[1] Künstle, D.-E., von Luxburg, U., & Wichmann, F. A. (2022a). Estimating the perceived dimension of psychophysical stimuli using triplet accuracy and hypothesis testing. *Journal of Vision*, *22*(13), 5

A.1 Variations of similarity judgment tasks

All experimental tasks used for psychophysical scaling measure the relative similarity between stimulus pairs, either directly as in rating experiments or indirectly as in JND-based or comparison-based methods used in the work at hand (see main paper). Experiments with comparison-based methods ask observers to order the perceived similarity or dissimilarity of stimulus pairs in one way or another. The ordering is most apparent in the quadruplet task, which asks per trial which of two pairs is most similar (dissimilar), and in the triplet task, which asks for three stimuli if left or right is more similar (dissimilar) to the center. The triplet task can be generalized to n-choose-k and n-rank-k tasks presenting n stimuli in addition to the anchor and asking for the k of n most similar (dissimilar) stimuli and in n-rank-k tasks for their ranking. A response to these tasks can be used to calculate multiple triplets of the anchor, the chosen, and another presented stimulus. Different tasks without anchor stimuli ask for the odd-one-out or the most central of three presented stimuli. The pairwise comparisons are most dissimilar for the odd-one-out and most similar for the most-central stimulus.

Table A.1 summarises possible conversions from various comparison-based tasks to triplets, such that these responses are usable with our proposed procedure. These converted triplets are partially dependent such that scaling performance might not be comparable between sampled and converted numbers of triplets. From a psychological perspective, one should be cautious, however, when comparing responses from conversions, because the actual task, the instructions, and the context (i.e. fewer or more presented stimuli) probably influence the responses. Mathematically these conversions are sound as long as the triangle inequality holds in the perceptual space.

Table A.1: Conversions from other comparison-based tasks to triplets. Triplets denote the response by order of the stimulus indices (anchor, chosen, other), and curly brackets are a short notation for repetition of the same triplet with all index variants in the bracket, e.g. $(1, \{2,3\}, \{4,5\})$ means triplets $(1,2,4), (1,2,5), (1,3,4)$, and $(1,3,5)$. In the examples below, we denote duplicated triplets with exchanges in one position by curly braces.

Task	Presented	Chosen	Triplet mapping	Example of task
8-choose-2	$(i, j, \ldots, q)$	j, k	$(i, \{j,k\}, \{l, \ldots, q\})$	Roads and Mozer (2019)
8-rank-2	$(i, j, \ldots, q)$	j, k	prev. and (i, j, k)	Roads and Mozer (2019)
odd-one-out	(i, j, k)	k	(i, j, k) and (j, i, k)	Hebart et al. (2020)
quadruplet	$((i, j), (k, l))$	(i, j)	$(i, j, l$ or $k)$ if $i == k$ or l $(j, i, l$ or $k)$ if $j == k$ or l	Maloney and Yang (2003)

A.2 Normal distribution of accuracy samples

The procedure presented in chapter 3 assumes normally distributed test accuracies from repeated scale estimates, which is grounded in this paper's main part from a theoretical perspective. Here we show the practical grounding in the form of a brief simulation experiment and a statistical test for normality.

We looked at two different accuracy samples: Accuracies from 100 independently simulated triplet datasets of the same ground-truth scale to approximate the actual accuracy distribution, the so called *hand-off* accuracies, and cross-validation accuracies of a single simulated dataset such as is used in our procedure. We simulated every dataset of $2,000$ triplets with a 3D-normal scale (medium noise) as described in the paper's simulation section. In the hand-off setting, we estimated the scale with $1,800$ triplets and calculated the accuracy with 200, and in the cross-validation configuration, we used ten repetitions of 10-folds.

The histogram of both accuracy samples is shown in Figure A.1 along with the sample means (vertical line) and intervals of two standard deviations (dashed line). The hand-off accuracy means, used as a proxy of the actual accuracy, is included in the corrected interval of the cross-validation samples but overestimates the accuracy spread.

The normality of both accuracy samples was tested with a combined skew and kurtosis test (D'Agostino & Pearson, 1973). Both settings failed to reject the null hypothesis; their samples are normally distributed (CV: $s^2 + k^2 = 2.46, p > .05$, test set: $s^2 + k^2 = 0.47, p > .05$).

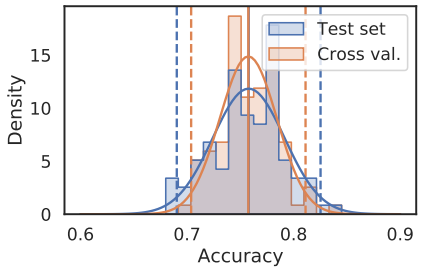


Figure A.1: Histogram of accuracy samples from independent test sets and cross-validation. Vertical lines show the mean and two standard deviations (dashed), and the line shows a corresponding normal distribution PDF. The cross-validation samples have a smaller deviation, which is the variation underestimated to be corrected in statistical tests (Nadeau & Bengio, 2003).

A.3 Noise visualization

The influence of triplet number and the judgment noise on scale estimates is illustrated in Figure A.2 by showing scale estimates from multiple noisy simulation runs. The coordinates were aligned in terms of rotation, scale, and translation with the Procrustes method (Gower, 1975).

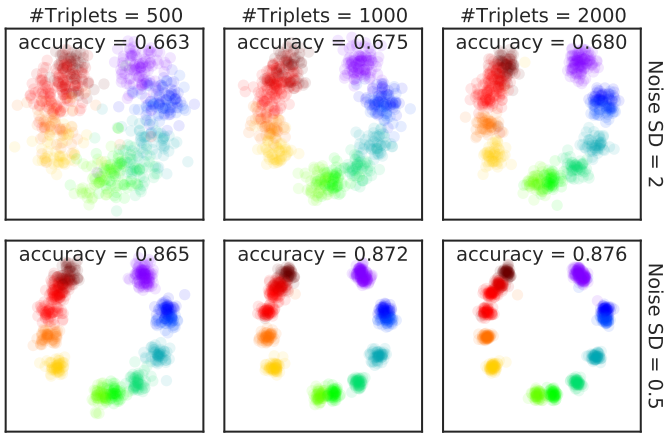


Figure A.2: The robustness of hue embeddings increases with the triplet number and decreases with the judgment noise.

A.4 Algorithm details

The following pseudo-code shows the algorithmic details of repeated cross-validation and the testing corrections used in our dimension testing procedure.

---

function REPEATEDCVSCORES(EMBEDDING, SCORE, $\mathcal{T}, r, k$)
  for all $j \in 1..r$ **do** ▷ Repeat CV r times.
    $\mathcal{T}^* \leftarrow$ SHUFFLE($\mathcal{T}$)
    $\mathcal{S} \leftarrow$ CROSS_VAL_PARTITION($|\mathcal{T}|, k$)
    for all $(\boldsymbol{s}_{\text{train}}, \boldsymbol{s}_{\text{test}}) \in \mathcal{S}$ **do** ▷ k-fold Cross Validation.
      $X \leftarrow$ EMBEDDING(d, $\mathcal{T}^*[\boldsymbol{s}_{\text{train}}]$)
      $\boldsymbol{u}_{i,j} \leftarrow$ SCORE(X, $\mathcal{T}^*[\boldsymbol{s}_{\text{test}}]$) ▷ Evaluate on test triplets.
    end for
  end for
  return $\boldsymbol{u}$ ▷ $r \cdot k$ test scores.
end function

function CORRECTEDTTEST($\boldsymbol{d}, n_{\text{train}}, n_{\text{train}}$)
  $\sigma_{\bar{d}} \leftarrow \sqrt{\frac{1}{|\boldsymbol{d}|} + \frac{n_{\text{test}}}{n_{\text{train}}}} \cdot \text{STD}(\boldsymbol{d})$ ▷ Correction of Nadeau and Bengio (2003).
  $t \leftarrow \frac{\text{MEAN}(\boldsymbol{d})}{\sigma_{\bar{d}}}$ ▷ t-test statistic.
  return STUDENT_T_PDF(t, $df \leftarrow |\boldsymbol{d}| - 1$)
end function

function HOLMMULTITESTCORRECTION($p_1, \ldots, p_k$)
  ▷ Holm (1979).
  $V \leftarrow \{1, \ldots, k\}$
  for $i \leftarrow k$ to 1 **do**
    $j \leftarrow \arg\min_{d \in V} p_d$
    $p^*_d \leftarrow \frac{p_d}{i}$
    $V \leftarrow V \setminus \{d\}$
  end for
  return $p^*_1, \ldots, p^*_k$
end function

A.5 Simulating a psychophysical experiment

Here we show simulation result where the ground-truth scale is inspired by actual psychophysical scales— the idealized hue and pitch perception as a wheel and a helix.

A.5.1 The hue perception wheel

This experiment used a two-dimensional hue circle as a realistic example of multi-dimensional ground-truth scales (Figure 3.1). While we simulate triplets from this ground-truth scale, the ground-truth scale is not artificial but estimated from psychological data. The scale was estimated from pairwise hue dissimilarities (Ekman, 1954) with the multi-dimensional scaling algorithm (Shepard, 1962).

Our procedure correctly estimated two dimensions in most settings, as shown in Figure A.3. Just two high noise simulations underestimated the dimension, consistent with the other simulation experiments.

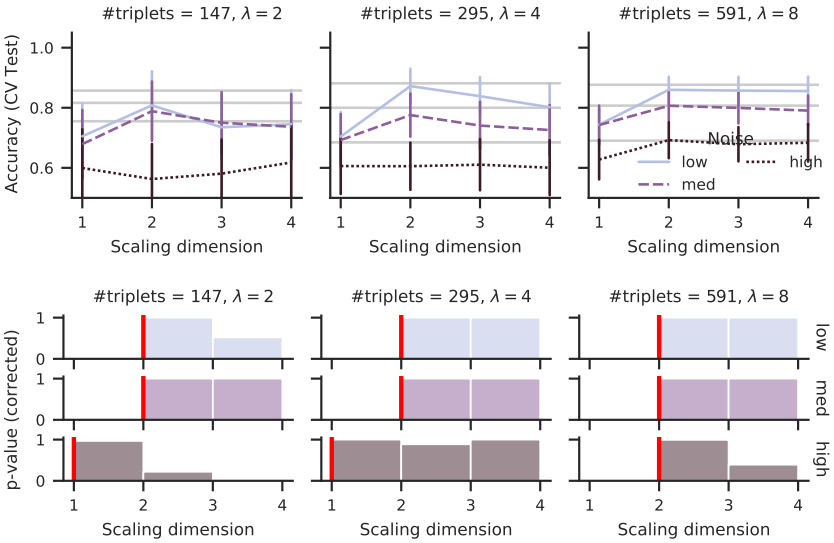


Figure A.3: Triplet accuracy and p-values for embedding triplets of a color wheel to a different dimension. The intrinsic dimension of two is correctly identified (vertical colored line).

A.5.2 The pitch perception helix

The ground-truth scale used in this section is a three-dimensional helix (Figure A.4), that is inspired from models of pitch perception (Shepard, 1965) but not based on behavioral data. As in the simulation experiments, we created a ground-truth scale and simulated responses, including normally distributed judgment noise. The ground-truth helix has three rotations with 12 tones (an octave) each, where the height of a rotation equals the helix's diameter.

Our procedure reconstructs the three-dimensional structure in the setting with high noise and low accuracy (Figure A.5). Surprisingly, the noisier setting shows another reasonably accurate representation with a single dimension, an unrolled version of the helix. This more straightforward, unrolled representation is preferred if less data is

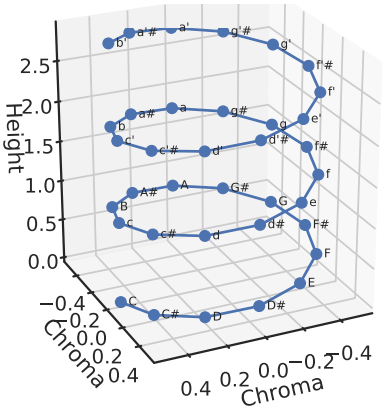


Figure A.4: Pitch helix

available. This trade-off between unrolled 1D and helix 3D representation should depend on the helix's diameter-to-height ratio.

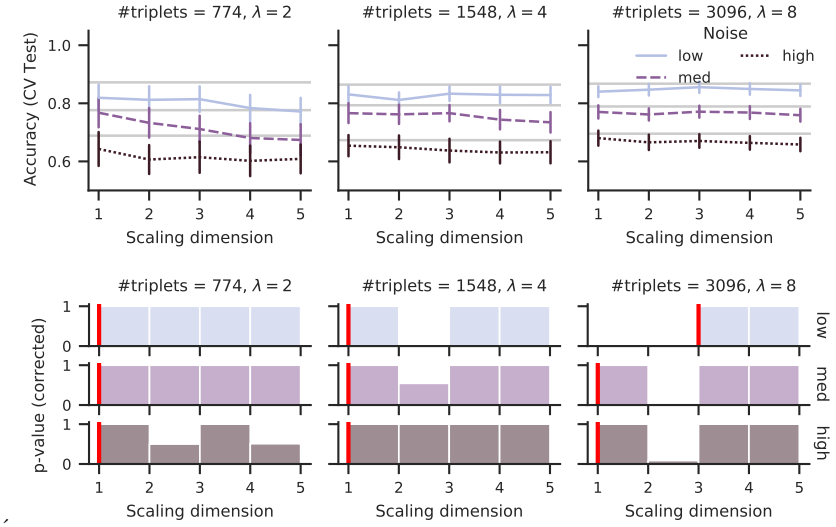


Figure A.5: Triplet accuracy and p-values for reconstructing a simulated pitch helix with different judgment noise (colors) and dataset size. The p-values tell two different interpretations: From noisy or small datasets, just the one-dimensional perceived pitch is reconstructed. p-values for the large dataset show, that–provided enough data–a three dimensional representation can represent additional nuances (the helix-like similarities between octaves).

A.6 Overview of simulation results

The Figure A.6 shows an overview of dimension estimates across all the *normal* datasets. No simulation overestimated the ground-truth dimension. While most simulations predicted the correct dimensionality, some underestimated it, especially at high noise and large ground-truth dimensions. The noise might shadow distinctive distance information of additional dimensions, so we can interpret these dimension estimates as a lower-bound dimension estimate. In psychological practice, such conservative or lower-bound dimension estimates are beneficial as they provide the simplest model that explains the collected data—given the inherent noise in the data.

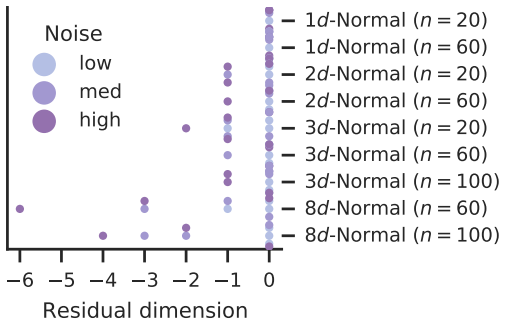


Figure A.6: Overview of difference between estimated and ground-truth dimension. Overestimating dimension occurred just for one-dimensional datasets, while underestimation occurred more often for higher dimensions and larger noise.

A.6.1 *Detailed simulation results*

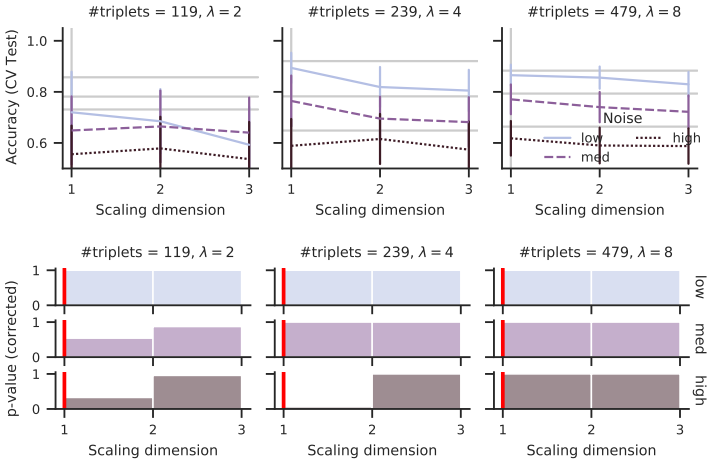


Figure A.7: 1D-normal, $n = 20$

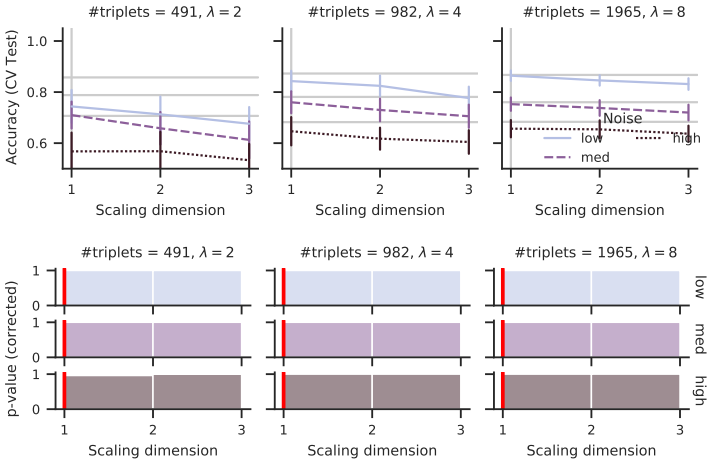


Figure A.8: 1D-normal, $n = 60$

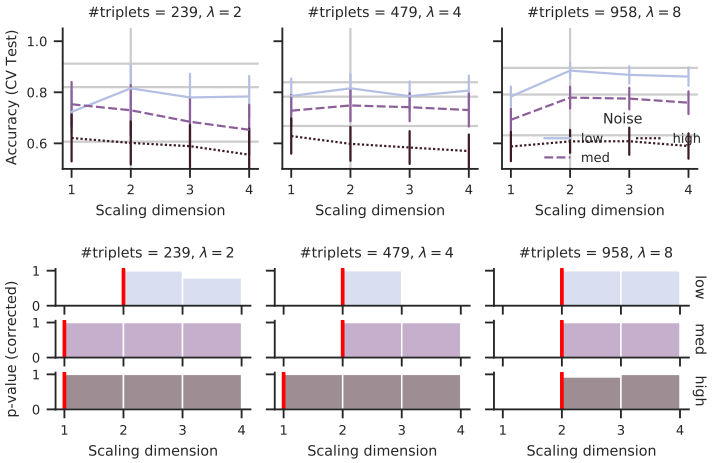


Figure A.9: 2D-normal, $n = 20$

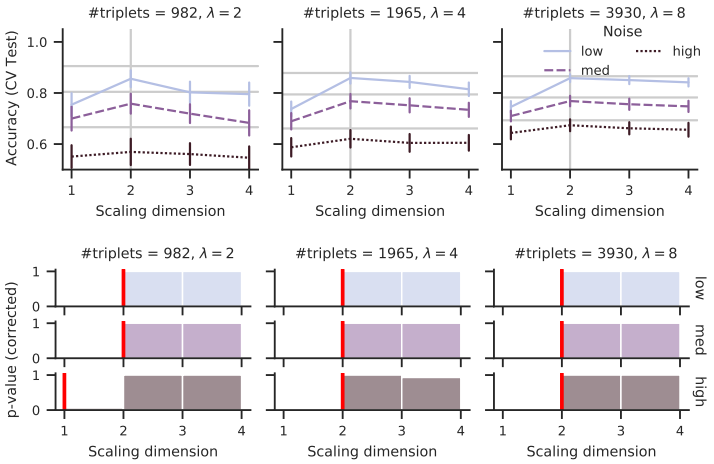


Figure A.10: 2D-normal, $n = 60$

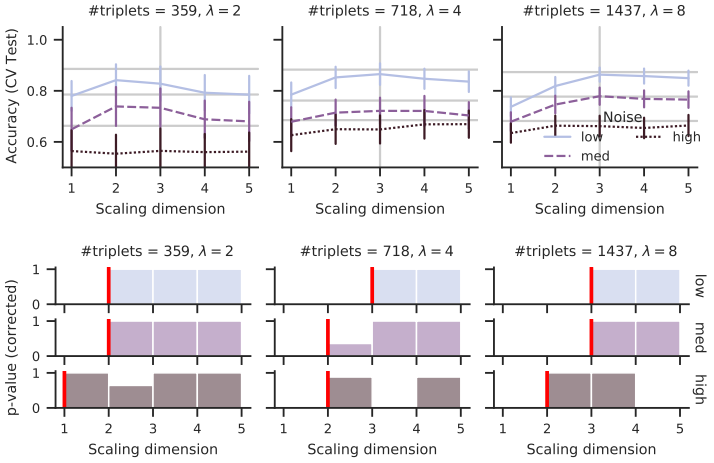


Figure A.11: 3D-normal, $n = 20$

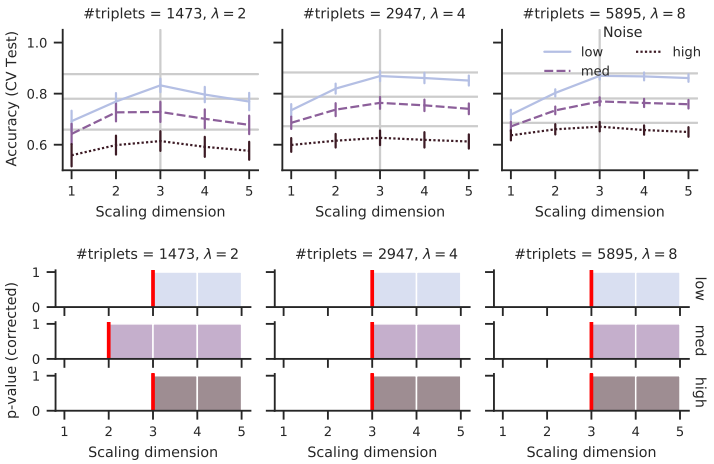


Figure A.12: 3D-normal, $n = 60$

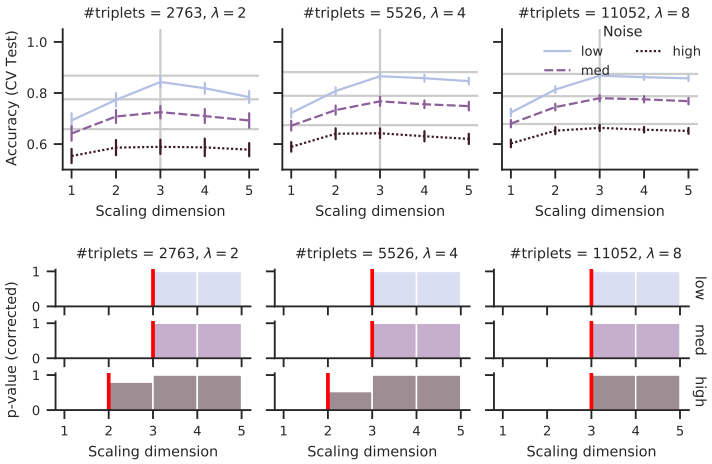


Figure A.13: 3D-normal,$n = 100$

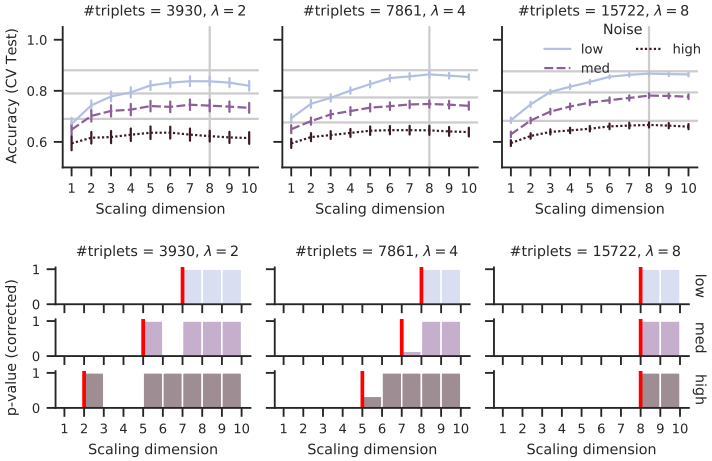


Figure A.14: 8D-normal, $n = 60$

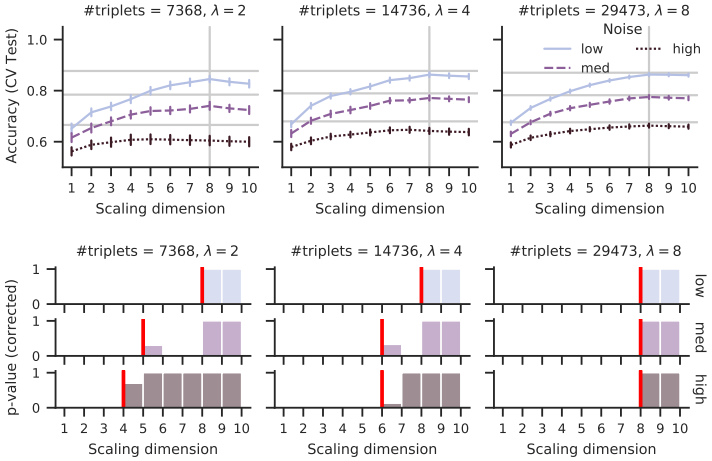


Figure A.15: 8D-normal,$n = 100$

B Supplementary lens distortion material

The content of this chapter equals the supplementary of our published article[1]. The author contributions are described in chapter 5.

[1] Sauer, Y., Künstle, D.-E., Wichmann, F. A., & Wahl, S. (2024). An objective measurement approach to quantify the perceived distortions of spectacle lenses. *Scientific Reports*, *14*(1), 3967

B.1 Individual subject data

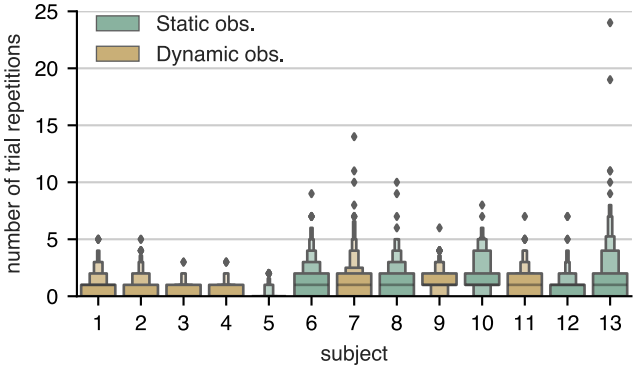


Figure B.1: Individual distributions of number of repeated trials per subject.

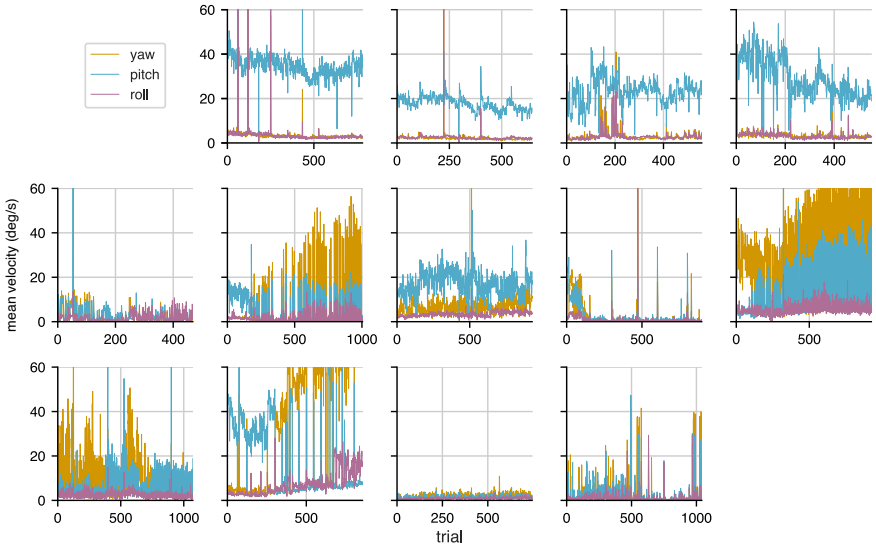


Figure B.2: Individual head-movement components for each subject over all trials.

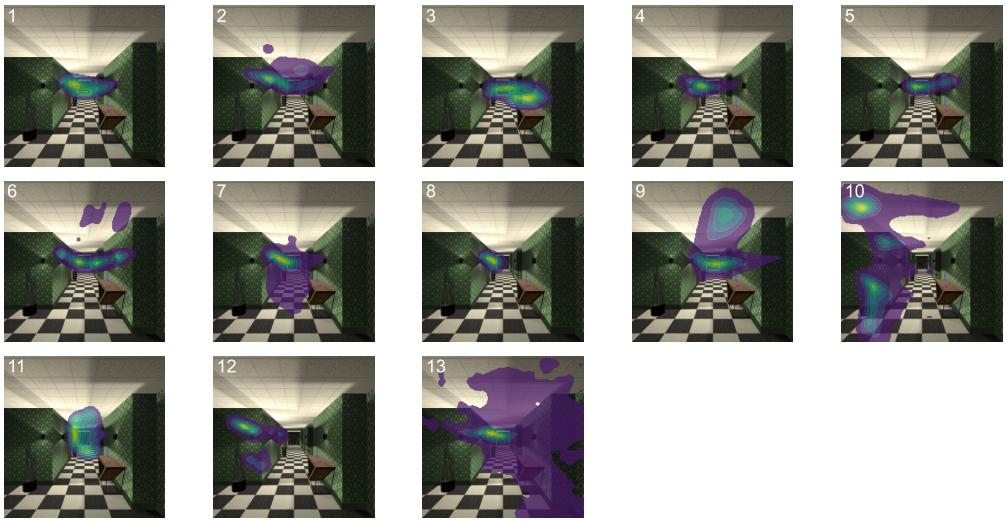


Figure B.3: Individual gaze distribution in the scene. The gaze target in the 3D scene is determined by head position and gaze direction. During runtime, the intersection point of the binocular gaze vector and the 3D environment was calculated using a ray cast originating at the headset position. Those 3D points were recorded in addition to all gaze vector samples. From the 3D points, we can analyze the 3D gaze distribution in the scene. For visualization, an image of the scene from a fixed viewpoint (the seating position of subjects in the scene) is overlayed with the 3D gaze samples mapped into the scene for the defined viewpoint and smoothed using a Gaussian kernel.

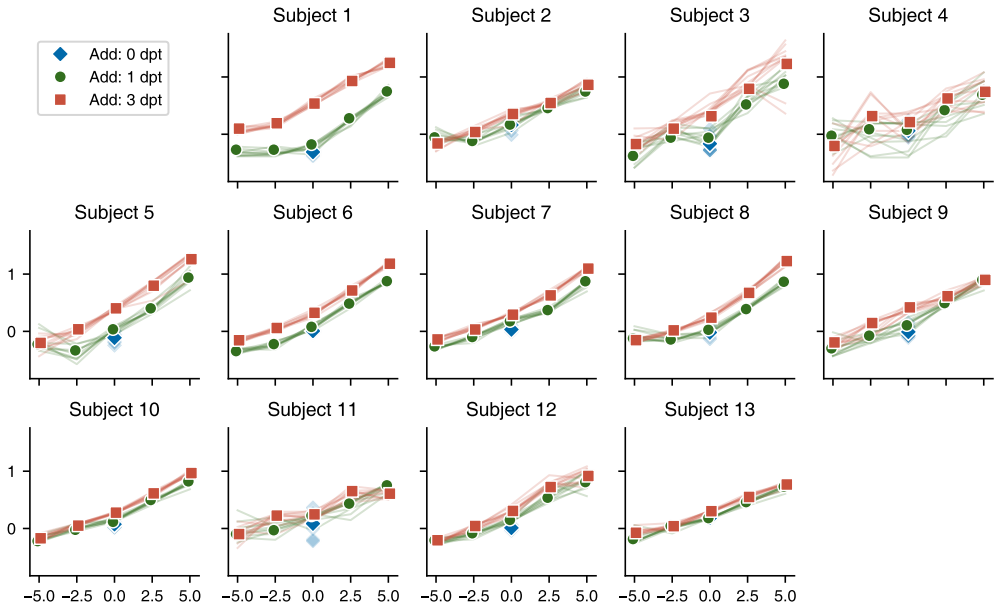


Figure B.4: Scales per subject (opaque markers) with bootstrapped variants (transparent markers and lines). The bootstrapped variants are scale estimates based not on all triplets but a random subset of 95% to increase variability. These resamples ought to test the stability of the estimate by simulating how much the scale estimates would change if different triplet questions were asked. For most subjects, these bootstrapped scales are very close to the scale of all triplets, indicating that collecting additional trials would not substantially change the estimate.

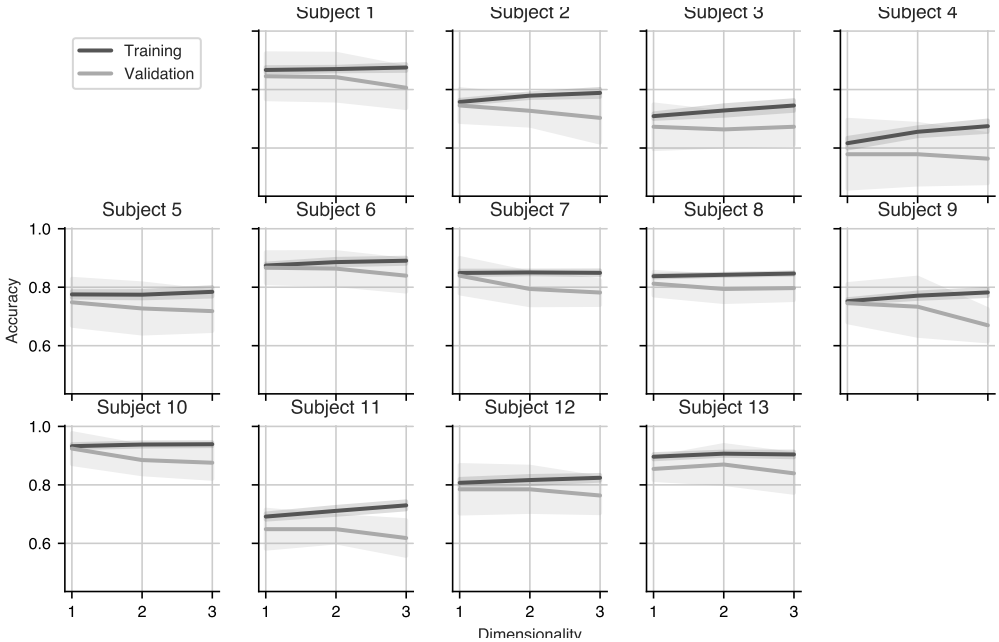


Figure B.5: Embedding accuracy depending on embedding dimensionality. Accuracy is the proportion of trials where the response can be predicted from the scale, which was fitted with different dimensionality. The embedding accuracies for all subjects do not increase with the dimensionality, which indicates that a 1D scale already fits the responses sufficiently.
The line and bands show the accuracy mean and standard deviation of 10 so-called cross-validation splits, where the scale was fitted on 90% of the trials ("Training") and validated on the remaining 10% trials. The accuracy of the validation trials ought to approximate the predictive performance for unseen responses. Additionally, the validation accuracy can be understood as an indicator of how consistently subjects respond—subjects 4 and 11 show more inconsistencies than the others.

B.2 Stimuli

Figure B.6: Virtual indoor environment used for the psychophysical experiment in the game engine Unity.

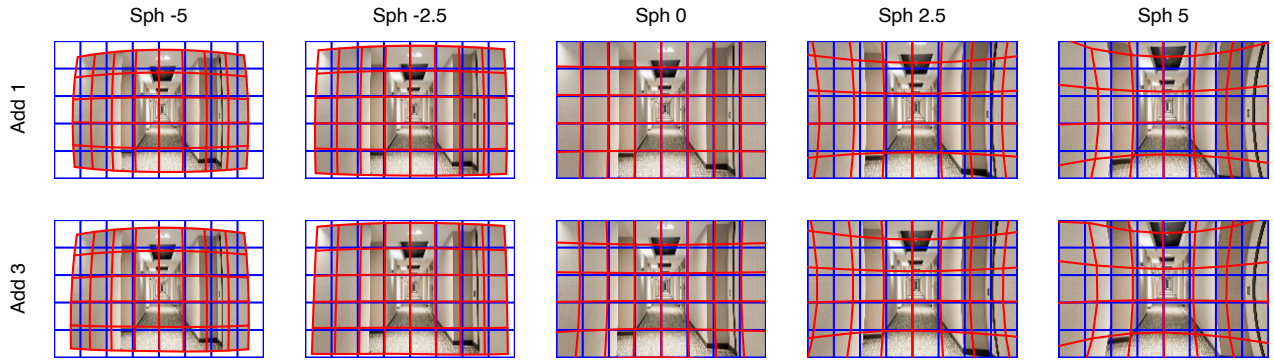


Figure B.7: Overview of all ten distortions used in the experiment. The far correction power *Sph* varied between −5 dpt to 5 dpt and the additional power for near vision was 1 dpt or 3 dpt.

Bibliography

Agarwal, S., Wills, J., Cayton, L., Lanckriet, G., Kriegman, D., & Belongie, S. (2007). Generalized non-metric multidimensional scaling. *Artificial Intelligence and Statistics*, 11–18.

Aguilar, G., & Maertens, M. (2020). Toward reliable measurements of perceptual scales in multiple contexts. *Journal of Vision*, *20*(4), 19.

Aguilar, G., Wichmann, F. A., & Maertens, M. (2017). Comparing sensitivity estimates from MLDS and forced-choice methods in a slant-from-texture experiment. *Journal of Vision*, *17*(1), 37.

Ahmed, S., Mula, R. S., & Dhavala, S. S. (2020, January 1). A framework for democratizing AI.

Alvarez, T. L., Han, S., Kania, C., Kim, E., Tsang, O., Semmlow, J. L., Granger-Donetti, B., & Pedrono, C. (2009). Adaptation to progressive lenses by presbyopes. *2009 4th international IEEE/EMBS conference on neural engineering*, 143–146.

Amid, E., & Ukkonen, A. (2015, July 7). Multiview triplet embedding: Learning attributes in multiple maps. In F. Bach & D. Blei (Eds.), *Proceedings of the 32nd international conference on machine learning* (pp. 1472–1480, Vol. 37). PMLR.

Anderton, J., & Aslam, J. (2019). Scaling up ordinal embedding: A landmark approach. *International conference on machine learning*, 282–290.

Ansel, J., Yang, E., He, H., Gimelshein, N., Jain, A., Voznesensky, M., Bao, B., Bell, P., Berard, D., Burovski, E., Chauhan, G., Chourdia, A., Constable, W., Desmaison, A., DeVito, Z., Ellison, E., Feng, W., Gong, J., Gschwind, M., … Chintala, S. (2024). PyTorch 2: Faster machine learning through dynamic python bytecode transformation and graph compilation. *Proceedings of the 29th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 2*, 929–947.

Arias-Castro, E. (2017). Some theory for ordinal embedding. *Bernoulli*, *23*(3), 1663–1693.

Attneave, F. (1950). Dimensions of similarity. *The American Journal of Psychology*, *63*(4), 516–556.

Balcan, M.-F., Vitercik, E., & White, C. (2016). Learning combinatorial functions from pairwise comparisons. *Conference on learning theory*, 310–335.

Barbero, S., & Portilla, J. (2017). Simulating real-world scenes viewed through ophthalmic lenses. *Journal of the Optical Society of America A, Optics and Image Science*, *34*(8), 1301–1308.

Bartleson, C. J. (1976). Brown*. *Color Research & Application*, *1*(4), 181–191.

Bauer, P., & Budde, M. (1994). Multiple testing for detecting efficient dose steps. *Biometrical Journal*, *36*(1), 1–15.

Beck, J., & Shaw, W. A. (1965). Magnitude of the standard, numerical value of the standard, and stimulus spacing in the estimation of loudness. *Perceptual and Motor Skills*, *21*(1), 151–156.

Benjamin, D. J., Berger, J. O., Johannesson, M., Nosek, B. A., Wagenmakers, E.-J., Berk, R., Bollen, K. A., Brembs, B., Brown, L., Camerer, C., Cesarini, D., Chambers, C. D., Clyde, M., Cook, T. D., De Boeck, P., Dienes, Z., Dreber, A., Easwaran, K., Efferson, C., … Johnson, V. E. (2018). Redefine statistical significance. *Nature Human Behaviour*, *2*(1), 6–10.

Bimler, D., Kirkland, J., & Jacobs, R. (2000). Colour-vision tests considered as a special case of multidimensional scaling. *Color Research & Application*, *25*(3), 160–169.

Bingham, E., Chen, J. P., Jankowiak, M., Obermeyer, F., Pradhan, N., Karaletsos, T., Singh, R., Szerlip, P., Horsfall, P., & Goodman, N. D. (2019). Pyro: Deep universal probabilistic programming. *Journal of Machine Learning Research*, *20*(28), 1–6.

Block, A., Jia, Z., Polyanskiy, Y., & Rakhlin, A. (2021). Intrinsic dimension estimation using wasserstein distances.

Bonferroni, C. (1936). Teoria statistica delle classi e calcolo delle probabilita. *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commericiali di Firenze*, *8*, 3–62.

Bonnardel, V., Beniwal, S., Dubey, N., Pande, M., Knoblauch, K., & Bimler, D. (2016). Perceptual color spacing derived from maximum likelihood multidimensional scaling. *Journal of the Optical Society of America A*, *33*(3), A30.

Borg, I., & Groenen, P. J. (2005). *Modern multidimensional scaling: Theory and applications*. Springer Science & Business Media.

Bosten, J. M., & Boehm, A. E. (2014). Empirical evidence for unique hues? *JOSA A*, *31*(4), A385–A393.

Bouckaert, R. R., & Frank, E. (2004). Evaluating the replicability of significance tests for comparing learning algorithms. In T. Kanade, J. Kittler, J. M. Kleinberg, F. Mattern, J. C. Mitchell, O. Nier-

strasz, C. Pandu Rangan, B. Steffen, M. Sudan, D. Terzopoulos, D. Tygar, M. Y. Vardi, G. Weikum, H. Dai, R. Srikant, & C. Zhang (Eds.), *Advances in knowledge discovery and data mining* (pp. 3–12, Vol. 3056). Springer Berlin Heidelberg.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32.

Brown, A. M., Lindsey, D. T., & Guckes, K. M. (2011). Color names, color categories, and color-cued visual search: Sometimes, color perception is not categorical. *Journal of Vision*, *11*(12), 2.

Brunswik, E. (1955). Representative design and probabilistic theory in a functional psychology. *Psychological Review*, *62*(3), 193–217.

Budde, M., & Bauer, P. (1989). Multiple test procedures in clinical dose finding studies. *Journal of the American Statistical Association*, *84*(407), 792–796.

Buitinck, L., Louppe, G., Blondel, M., Pedregosa, F., Mueller, A., Grisel, O., Niculae, V., Prettenhofer, P., Gramfort, A., Grobler, J., Layton, R., Vanderplas, J., Joly, A., Holt, B., & Varoquaux, G. (2013). API design for machine learning software: Experiences from the scikit-learn project. *arXiv:1309.0238 [cs.LG]*.

Burley, B., & Studios, W. D. A. (2012). Physically-based shading at disney. *ACM SIGGRAPH*, *2012*, 1–7.

Camastra, F., & Staiano, A. (2016). Intrinsic dimension estimation: Advances and open problems. *Information Sciences*, *328*, 26–41.

Caragiannis, I. (2017). Recent advances in large-scale peer grading. In U. Endriss (Ed.), *Computational social choice* (pp. 327–344). AI Access.

Carroll, J. D., & Chang, J.-J. (1970). Analysis of individual differences in multidimensional scaling via an n-way generalization of "eckart-young" decomposition. *Psychometrika*, *35*(3), 283–319.

Chandler, D. M. (2013). Seven challenges in image quality assessment: Past, present, and future research. *ISRN Signal Processing*, *2013*, 1–53.

Charman, W. N. (2014). Developments in the correction of presbyopia i: Spectacle and contact lenses. *Ophthalmic and Physiological Optics*, *34*(1), 8–29.

Charrier, C., Maloney, L. T., Cherifi, H., & Knoblauch, K. (2007). Maximum likelihood difference scaling of image quality in compression-degraded images. *Journal of the Optical Society of America A*, *24*(11), 3418.

Chen, T., Kornblith, S., Norouzi, M., & Hinton, G. (2020). A simple framework for contrastive learning of visual representations. *arXiv preprint arXiv:2002.05709*.

Chen, T., Kornblith, S., Swersky, K., Norouzi, M., & Hinton, G. (2020). Big self-supervised models are strong semi-supervised learners. *arXiv preprint arXiv:2006.10029*.

D'Agostino, R., & Pearson, E. S. (1973). Tests for departure from normality. empirical results for the distributions of b2 and b1. *Biometrika*, *60*(3), 613–622.

Davison, M. L., & Sireci, S. G. (2000, January 1). Multidimensional scaling. In H. E. A. Tinsley & S. D. Brown (Eds.), *Handbook of applied multivariate statistics and mathematical modeling* (pp. 323–352). Academic Press.

Demiralp, Ç., Bernstein, M. S., & Heer, J. (2014). Learning perceptual kernels for visualization design. *IEEE Transactions on Visualization and Computer Graphics*, *20*(12), 1933–1942.

de Vazelhes, W., Carey, C., Tang, Y., Vauquier, N., & Bellet, A. (2020). Metric-learn: Metric learning algorithms in python. *Journal of Machine Learning Research*, *21*(138), 1–6.

Devinck, F., & Knoblauch, K. (2012). A common signal detection model accounts for both perception and discrimination of the watercolor effect. *Journal of Vision*, *12*(3), 19–19.

Dietterich, T. G. (1998). Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Computation*, *10*(7), 1895–1923.

Efron, B. (1987). Better bootstrap confidence intervals. *Journal of the American Statistical Association*, *82*(397), 171–185.

Egan, J. P. (1975). *Signal detection theory and ROC-analysis*. Academic Press.

Ekman, G. (1954). Dimensions of color vision. *The Journal of Psychology*, *38*(2), 467–474.

Fairchild, M. D. (2013). Color appearance models. In *Color appearance models* (pp. 199–212). John Wiley & Sons, Ltd.

Fechner, G. T. (1860). *Elemente der psychophysik*. Breitkopf & Härtel.

Fleming, R. W. (2017). Material perception. *Annual Review of Vision Science*, *3*(1), 365–388.

Fleming, R. W., Jäkel, F., & Maloney, L. T. (2011). Visual perception of thick transparent materials. *Psychological Science*, *22*(6), 812–820.

Fletcher, R. (1987). *Practical methods of optimization; (2nd ed.)* Wiley-Interscience.

Freiesleben, T., König, G., Molnar, C., & Tejero-Cantero, A. (2022, November 15). Scientific inference with interpretable machine learning: Analyzing models to learn about real-world phenomena.

Fricke, T. R., Tahhan, N., Resnikoff, S., Papas, E., Burnett, A., Ho, S. M., Naduvilath, T., & Naidoo, K. S. (2018). Global prevalence of presbyopia and vision impairment from uncorrected presbyopia: Systematic review, meta-analysis, and modelling. *Ophthalmology*, *125*(10), 1492–1499.

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, *29*(5), 1189–1232.

Gescheider, G. A. (1988). Psychophysical scaling. *Annual Review of Psychology*, *39*(1), 169–200.

Gescheider, G. A. (2013a). The measurement of sensory attributes and discrimination scales. In *Psychophysics: The fundamentals* (pp. 183–206). Taylor & Francis.

Gescheider, G. A. (2013b). Partition scales. In *Psychophysics: The fundamentals* (pp. 183–206). Taylor & Francis.

Ghosh, N., Chen, Y., & Yue, Y. (2019). Landmark ordinal embedding. In H. Wallach, H. Larochelle, A. Beygelzimer, F. dAlché-Buc, E. Fox, & R. Garnett (Eds.), *Advances in neural information processing systems* (Vol. 32). Curran Associates, Inc.

Ghoshdastidar, D., Perrot, M., & von Luxburg, U. (2019). Foundations of comparison-based hierarchical clustering. *Advances in Neural Information Processing Systems*, *32*.

Goldstein, E. B. (2007). Sound, the auditory sytem, and pitch perception. In *Sensation and perception* (7th ed., pp. 233–264). Thomson Wadsworth.

Goldstone, R. (1994). An efficient method for obtaining similarity data. *Behavior Research Methods, Instruments, & Computers*, *26*(4), 381–386.

Gorinova, M., Moore, D., & Hoffman, M. (2020). Automatic reparameterisation of probabilistic programs. *Proceedings of the 37th International Conference on Machine Learning*, 3648–3657.

Gower, J. C. (1975). Generalized procrustes analysis. *Psychometrika*, *40*(1), 33–51.

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics*. John Wiley.

Gronau, Q. F., & Lee, M. D. (2020). Bayesian inference for multidimensional scaling representations with psychologically interpretable metrics. *Computational Brain & Behavior*, *3*(3), 322–340.

Haghiri, S., Garreau, D., & von Luxburg, U. (2018). Comparison-based random forests. *Proceedings of the 35th International Conference on Machine Learning (ICML)*.

Haghiri, S., Ghoshdastidar, D., & von Luxburg, U. v. (2017). Comparison-based nearest neighbor search. In A. Singh & J. Zhu (Eds.), *Proceedings of the 20th international conference on artificial intelligence and statistics* (pp. 851–859, Vol. 54). PMLR.

Haghiri, S., Rubisch, P., Geirhos, R., Wichmann, F., & von Luxburg, U. (2019). Comparison-based framework for psychophysics: Lab versus crowdsourcing. *arXiv:1905.07234 [cs, stat]*.

Haghiri, S., Wichmann, F. A., & von Luxburg, U. (2020). Estimation of perceptual scales using ordinal embedding. *Journal of Vision*, *20*(9), 14.

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M. H., Brett, M., Haldane, A., del Río, J. F., Wiebe, M., Peterson, P., ... Oliphant, T. E. (2020). Array programming with NumPy. *Nature*, *585*(7825), 357–362.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). Model assessment and selection. In *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed., p. 763). Springer-Verlag.

Head, M. L., Holman, L., Lanfear, R., Kahn, A. T., & Jennions, M. D. (2015). The extent and consequences of p-hacking in science. *PLOS Biology*, *13*(3), e1002106.

Hebart, M. N., Dickter, A. H., Kidder, A., Kwok, W. Y., Corriveau, A., Wicklin, C. V., & Baker, C. I. (2019). THINGS: A database of 1,854 object concepts and more than 26,000 naturalistic object images. *PLOS ONE*, *14*(10), e0223792.

Hebart, M. N., Zheng, C. Y., Pereira, F., & Baker, C. I. (2020). Revealing the multidimensional mental representations of natural objects underlying human similarity judgements. *Nature Human Behaviour*, *4*(11), 1173–1185.

Hefner, R. A. (1958). *Extensions of the law of comparative judgment to discriminable and multidimensional stimuli*. University of Michigan.

Heikinheimo, H., & Ukkonen, A. (2013). The crowd-median algorithm. *Proceedings of the AAAI conference on human computation and crowdsourcing*, *1*, 69–77.

Heim, E., Berger, M., Seversky, L., & Hauskrecht, M. (2015, November 6). Active perceptual similarity modeling with auxiliary information.

Ho, Y.-X., Landy, M. S., & Maloney, L. T. (2008). Conjoint measurement of gloss and surface texture. *Psychological science*, *19*(2), 196–204.

Hoffman, D. D., Singh, M., & Prakash, C. (2015). The interface theory of perception. *Psychonomic Bulletin & Review*, *22*(6), 1480–1506.

Hoffman, M. D., & Gelman, A. (2014). The no-u-turn sampler: Adaptively setting path lengths in hamiltonian monte carlo. *Journal of Machine Learning Research*, *15*(47), 1593–1623.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, *6*(2), 65–70.

Hölscher, J. (2022). *Perceptual reparametrization of image manipulation sliders* [Master Thesis]. Eberhard Karls Universität Tübingen.

Hooker, G. (2007). Generalized functional ANOVA diagnostics for high-dimensional functions of dependent variables. *Journal of Computational and Graphical Statistics*.

Huber, L. S., Künstle, D.-E., & Reuter, K. (2024). Tracing truth through conceptual scaling: Mapping people's understanding of abstract concepts.

Hutchings, N., Irving, E. L., Jung, N., Dowling, L. M., & Wells, K. A. (2007). Eye and head movement alterations in naïve progressive addition lens wearers. *Ophthalmic and Physiological Optics*, *27*(2), 142–153.

Hutter, F., Kotthoff, L., & Vanschoren, J. (Eds.). (2019). *Automated machine learning: Methods, systems, challenges*. Springer Nature.

Imrie, F., Cebere, B., McKinney, E. F., & Schaar, M. v. d. (2023). AutoPrognosis 2.0: Democratizing diagnostic and prognostic modeling in healthcare with automated machine learning. *PLOS Digital Health*, *2*(6), e0000276.

Jacoby, W. G., & Armstrong II, D. A. (2014). Bootstrap confidence regions for multidimensional scaling solutions. *American Journal of Political Science*, *58*(1), 264–278.

Jain, L., Jamieson, K. G., & Nowak, R. (2016). Finite sample prediction and recovery bounds for ordinal embedding. *Advances in Neural Information Processing Systems*, *29*.

Jäkel, F., Schölkopf, B., & Wichmann, F. A. (2008). Similarity, kernels, and the triangle inequality. *Journal of Mathematical Psychology*, *52*(5), 297–303.

Jalie, M. (2020). Modern spectacle lens design. *Clinical and Experimental Optometry*, *103*(1), 3–10.

Jamieson, K. G., Jain, L., Fernandez, C., Glattard, N. J., & Nowak, R. (2015). NEXT: A system for real-world development, evaluation, and application of active learning. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, & R. Garnett (Eds.), *Advances in neural information processing systems* (Vol. 28). Curran Associates, Inc.

Jamieson, K. G., & Nowak, R. D. (2011). Low-dimensional embedding using adaptively selected ordinal data. *2011 49th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, 1077–1084.

Johnson, L., Buckley, J. G., Scally, A. J., & Elliott, D. B. (2007). Multifocal spectacles increase variability in toe clearance and risk of tripping in the elderly. *Investigative ophthalmology & visual science*, *48*(4), 1466–1471.

Kaernbach, C. (1991). Poisson signal-detection theory: Link between threshold models and the gaussian assumption. *Perception & Psychophysics*, *50*(5), 498–506.

Kaernbach, C. (2001). Slope bias of psychometric functions derived from adaptive data. *Perception & Psychophysics*, *63*(8), 1389–1398.

Kiers, H. A. L., & Groenen, P. J. F. (2006). Visualizing dependence of bootstrap confidence intervals for methods yielding spatial configurations. In S. Zani, A. Cerioli, M. Riani, & M. Vichi (Eds.), *Data analysis, classification and the forward search* (pp. 119–126). Springer.

Kim, J., Marlow, P., & Anderson, B. L. (2011). The perception of gloss depends on highlight congruence with surface shading. *Journal of Vision*, *11*(9), 4.

Kingma, D. P., & Ba, J. (2015). Adam: A method for stochastic optimization. In Y. Bengio & Y. LeCun (Eds.), *3rd international conference on learning representations, ICLR 2015, san diego, CA, USA, may 7-9, 2015, conference track proceedings*.

Kleindessner, M., & von Luxburg, U. (2014). Uniqueness of ordinal embedding. *Proceedings of The 27th Conference on Learning Theory*, 40–67.

Kleindessner, M., & von Luxburg, U. (2015). Dimensionality estimation without distances. *Artificial Intelligence and Statistics*, 471–479.

Kleindessner, M., & von Luxburg, U. (2017). Lens depth function and k-relative neighborhood graph: Versatile tools for ordinal data analysis. *Journal of Machine Learning Research*, *18*(58), 1–52.

Knoblauch, K., & Maloney, L. T. (2008). MLDS: Maximum likelihood difference scaling in r. *Journal of Statistical Software*, *25*, 1–26.

Knoblauch, K., & Maloney, L. T. (2012a). Maximum likelihood difference scaling. In *Modeling psychophysical data in r* (pp. 195–228). Springer New York.

Knoblauch, K., & Maloney, L. T. (2012b). *Modeling psychophysical data in r*. Springer New York.

Knoblauch, K., Marsh-Armstrong, B., & Werner, J. S. (2020). Suprathreshold contrast response in normal and anomalous trichromats. *JOSA A*, *37*(4), A133–A144.

Koenderink, J., Valsecchi, M., van Doorn, A., Wagemans, J., & Gegenfurtner, K. (2017). Eidolons: Novel stimuli for vision research. *Journal of Vision*, *17*(2), 7–7.

Krantz, D. H., Suppes, P., & Luce, R. D. (2006). *Additive and polynomial representations* (Vol. 1). Courier Corporation.

Kruskal, J. B. (1964a). Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, *29*(1), 1–27.

Kruskal, J. B. (1964b). Nonmetric multidimensional scaling: A numerical method. *Psychometrika*, *29*(2), 115–129.

Künstle, D.-E., & von Luxburg, U. (2024). Cblearn: Comparison-based machine learning in python. *Journal of Open Source Software*, *9*(98), 6139.

Künstle, D.-E., von Luxburg, U., & Wichmann, F. A. (2022a). Estimating the perceived dimension of psychophysical stimuli using triplet accuracy and hypothesis testing. *Journal of Vision*, *22*(13), 5.

Künstle, D.-E., von Luxburg, U., & Wichmann, F. A. (2022b). Estimating the perceived dimensionality of psychophysical stimuli using a triplet accuracy and hypothesis testing procedure. *Journal of Vision*, *22*(14), 3331.

Künstle, D.-E., & Wichmann, F. A. (2023). Measuring lightness constancy with varying realism. *Journal of Vision*, *23*(9), 5281.

Lagunas, M., Malpica, S., Serrano, A., Garces, E., Gutierrez, D., & Masia, B. (2019). A similarity measure for material appearance. *ACM Transactions on Graphics*, *38*(4), 1–12.

Letocha, C. E. (1990). The invention and early manufacture of bifocals. *Survey of ophthalmology*, *35*(3), 226–235.

Li, L., Malave, V. L., Song, A., & Angela, J. Y. (2016). Extracting human face similarity judgments: Pairs or triplets? *CogSci*, 1427–1432.

Liu, E. Y., Guo, Z., Zhang, X., Jojic, V., & Wang, W. (2012). Metric learning from relative comparisons by minimizing squared residual. *ICDM*, 978–983.

Logvinenko, A. D., & Maloney, L. T. (2006). The proximity structure of achromatic surface colors and the impossibility of asymmetric lightness matching. *Perception & Psychophysics*, *68*(1), 76–83.

Lohaus, M., Hennig, P., & von Luxburg, U. (2019). Uncertainty estimates for ordinal embeddings. *arXiv:1906.11655 [cs, stat]*.

Love, B. C., & Roads, B. D. (2021). Similarity as a window on the dimensions of object representation. *Trends in Cognitive Sciences*, *25*(2), 94–96.

Luce, R. D., & Tukey, J. W. (1964). Simultaneous conjoint measurement: A new type of fundamental measurement. *Journal of Mathematical Psychology*, *1*(1), 1–27.

Maloney, L. T., & Knoblauch, K. (2020). Measuring and modeling visual appearance. *Annual Review of Vision Science*, *6*(1), 519–537.

Maloney, L. T., & Yang, J. N. (2003). Maximum likelihood difference scaling. *Journal of Vision*, *3*(8), 5–5.

Mandal, A., Perrot, M., & Ghoshdastidar, D. (2023, April). A revenue function for comparison-based hierarchical clustering.

Marin, G., Terrenoire, E., & Hernandez, M. (2008). Compared distortion effects between real and virtual ophthalmic lenses with a simulator. *Proceedings of the 2008 ACM symposium on virtual reality software and technology*, 271–272.

Marlow, P. J., Kim, J., & Anderson, B. L. (2011). The role of brightness and orientation congruence in the perception of surface gloss. *Journal of Vision*, *11*(9), 16.

Marlow, P. J., Kim, J., & Anderson, B. L. (2012). The perception and misperception of specular surface reflectance. *Current Biology*, *22*(20), 1909–1913.

Meister, D. J., & Fisher, S. W. (2008). Progress in the spectacle correction of presbyopia. part 1: Design and development of progressive lenses. *Clinical and experimental optometry*, *91*(3), 240–250.

Men, H., Lin, H., Jenadeleh, M., & Saupe, D. (2021). Subjective image quality assessment with boosted triplet comparisons. *IEEE Access*, *9*, 138939–138975.

Merkel, J. (1888). Die abhangigkeit zwischen reiz und empfindung, erste abtheilung. *Philosophische Studien*, *4*, 541–594.

Metzen, H. M. (2016). *Jmetzen/gp_extras code repository*.

Miller, G. A., & Nicely, P. E. (1955). An analysis of perceptual confusions among some english consonants. *The Journal of the Acoustical Society of America*, *27*(2), 338–352.

Minkwitz, G. (1963). Über den flächenastigmatismus bei gewissen symmetrischen asphären. *Optica Acta: International Journal of Optics*, *10*(3), 223–227.

Muttenthaler, L., Zheng, C. Y., McClure, P., Vandermeulen, R. A., Hebart, M. N., & Pereira, F. (2022). VICE: Variational interpretable concept embeddings. *Advances in Neural Information Processing Systems*, *35*, 33661–33675.

Nadeau, C., & Bengio, Y. (2003). Inference for the generalization error. *Machine Learning*, *52*(3), 239–281.

Nießner, M., Sturm, R., & Greiner, G. (2012). Real-time simulation and visualization of human vision through eyeglasses on the GPU. *Proceedings of the 11th ACM SIGGRAPH international conference on virtual-reality continuum and its applications in industry*, 195–202.

Obein, G., Knoblauch, K., & Viéot, F. (2004). Difference scaling of gloss: Nonlinearity, binocularity, and constancy. *Journal of Vision*, *4*(9), 4.

Oh, M.-S., & Raftery, A. E. (2001). Bayesian multidimensional scaling and choice of dimension. *Journal of the American Statistical Association*, *96*(455), 1031–1044.

Pareto, V. (1896). *Cours d'économie politique* (Vol. 1). Librairie Droz.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., … Chintala, S. (2019). Py-

Torch: An imperative style, high-performance deep learning library. *Advances in Neural Information Processing Systems*, *32*.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine learning in python. *Journal of Machine Learning Research (JMLR)*, *12*(85), 2825–2830.

Perrot, M., Esser, P., & Ghoshdastidar, D. (2020). Near-optimal comparison based clustering. *Advances in Neural Information Processing Systems*, *33*, 19388–19399.

Perrot, M., Habrard, A., Muselet, D., & Sebban, M. (2014). Modeling perceptual color differences by local metric learning. In D. Fleet, T. Pajdla, B. Schiele, & T. Tuytelaars (Eds.), *Computer vision – ECCV 2014* (pp. 96–111). Springer International Publishing.

Phan, D., Pradhan, N., & Jankowiak, M. (2019, December 24). Composable effects for flexible and accelerated probabilistic programming in NumPyro.

Ponomarenko, N., Lukin, V., Zelensky, A., Egiazarian, K., Carli, M., & Battisti, F. (2009). TID2008-a database for evaluation of full-reference visual quality assessment metrics. *Advances of modern radioelectronics*, *10*(4), 30–45.

Pope, D. R. (2000). Progressive addition lenses: History, design, wearer satisfaction and trends. *Vision science and its applications*, NW9.

Poulton, E. C. (1968). The new psychophysics: Six models for magnitude estimation. *Psychological bulletin*, *69*(1), 1.

Pousset, N., Obein, G., & Razet, A. (2010). Visual experiment on LED lighting quality with color quality scale colored samples. *CIE 2010: Lighting Quality and Energy Efficiency*, 722–729.

Press, W. H., Teukolsky, S. A., Vetterling, W. T., & Flannery, B. P. (2007). *Numerical recipes 3rd edition: The art of scientific computing* (3rd ed.). Cambridge University Press.

Radonjić, A., Cottaris, N. P., & Brainard, D. H. (2019). The relative contribution of color and material in object selection. *PLoS computational biology*, *15*(4), e1006950.

Ramsay, J. O. (1969). Some statistical considerations in multidimensional scaling. *Psychometrika*, *34*(2), 167–182.

Ribe, R. (2022). Exploring psychophysical measurement in landscape aesthetics: Validity, reliability and signal detection via single- versus opposing-construct rating scales, with or without zeros. *Journal of Environmental Psychology*, *83*, 101862.

Rifai, K., & Wahl, S. (2016). Specific eye–head coordination enhances vision in progressive lens wearers. *Journal of Vision*, *16*(11), 5–5.

Roads, B. D., & Love, B. C. (2021). Enriching ImageNet with human similarity judgments and psychological embeddings. *Conference on Computer Vision and Pattern Recognition (CVPR)*.

Roads, B. D., & Love, B. C. (2024). Modeling similarity and psychological space. *Annual Review of Psychology*, *75*(1), annurev–psych–040323–115131.

Roads, B. D., & Mozer, M. C. (2019). Obtaining psychological embeddings through joint kernel and metric learning. *Behavior Research Methods*, *51*(5), 2180–2193.

Robinson, G. H. (1976). Biasing power law exponents by magnitude estimation instructions. *Perception & Psychophysics*, *19*(1), 80–84.

Rodríguez Celaya, J. A., Brunet Crosa, P., Ezquerra, N., & Palomar, J. (2005). A virtual reality approach to progressive lenses simulation. *XV congreso espanol de informatica grafica*.

Rogers, M., Knoblauch, K., & Franklin, A. (2016). Maximum likelihood conjoint measurement of lightness and chroma. *Journal of the Optical Society of America A*, *33*(3), A184.

Rojo, P., Royo, S., Ramírez, J., & Madariaga, I. (2014). Numerical implementation of generalized coddington equations for ophthalmic lens design. *Journal of Modern Optics*, *61*(3), 204–214.

Rosas, P., Wagemans, J., Ernst, M. O., & Wichmann, F. A. (2005). Texture and haptic cues in slant discrimination: Reliability-based cue weighting without statistically optimal cue combination. *JOSA A*, *22*(5), 801–809.

Rosas, P., Wichmann, F. A., & Wagemans, J. (2004). Some observations on the effects of slant and texture type on slant-from-texture. *Vision Research*, *44*(13), 1511–1535.

Rosas, P., Wichmann, F. A., & Wagemans, J. (2007). Texture and object motion in slant discrimination: Failure of reliability-based weighting of cues may be evidence for strong fusion. *Journal of Vision*, *7*(6), 3.

Roscher, R., Bohn, B., Duarte, M. F., & Garcke, J. (2020). Explainable machine learning for scientific insights and discoveries. *IEEE Access*, *8*, 42200–42216.

Ross, H. E. (1997). On the possible relations between discriminability and apparent magnitude. *British Journal of Mathematical and Statistical Psychology*, *50*(2), 187–203.

Sajjadi, M. S., Alamgir, M., & von Luxburg, U. (2016). Peer grading in a course on algorithms and data structures: Machine learning algorithms do not improve over simple baselines. *Proceedings of the Third (2016) ACM Conference on Learning @ Scale*, 369–378.

Samuel, A. L. (1959). Some studies in machine learning using the game of checkers. *IBM Journal of Research and Development*, *3*(3), 210–229.

Sattath, S., & Tversky, A. (1977). Additive similarity trees. *Psychometrika*, *42*(3), 319–345.

Sauer, Y., Künstle, D.-E., Wichmann, F., & Wahl, S. (2023a). Seeing the future of progressive glasses: A new perceptual approach to spectacle lens design.

Sauer, Y., Künstle, D.-E., Wichmann, F., & Wahl, S. (2023b). Psychophysical scale of optical distortions of multifocal spectacle lenses. *Journal of Vision*, *23*(9), 5215.

Sauer, Y., Künstle, D.-E., Wichmann, F. A., & Wahl, S. (2024). An objective measurement approach to quantify the perceived distortions of spectacle lenses. *Scientific Reports*, *14*(1), 3967.

Sauer, Y., Scherff, M., Lappe, M., Rifai, K., Stein, N., & Wahl, S. (2022). Self-motion illusions from distorted optic flow in multifocal glasses. *Iscience*, *25*(1).

Sauer, Y., Wahl, S., & Habtegiorgis, S. W. (2022). Realtime blur simulation of varifocal spectacle lenses in virtual reality. *SIGGRAPH asia 2022 technical communications*.

Sauer, Y., Wahl, S., & Rifai, K. (2020). Parallel adaptation to spatially distinct distortions. *Frontiers in Psychology*, *11*, 544867.

Schmid, A. C., & Anderson, B. L. (2017). Perceptual dimensions underlying lightness perception in homogeneous center-surround displays. *Journal of Vision*, *17*(2), 6.

Schönmann, I. (2021). *Similarity judgements of natural images: Instructions affect observers' decision criteria and consistency* (Bachelor thesis). University of Tübingen. Tübingen.

Schönmann, I., Künstle, D.-E., & Wichmann, F. A. (2022). Using an odd-one-out design affects consistency, agreement and decision criteria in similarity judgement tasks involving natural images. *Journal of Vision*, *22*(14), 3232.

Schroeder, H. W. (1984). Environmental perception rating scales: A case for simple methods of analysis. *Environment and Behavior*, *16*(5), 573–598.

Schütt, H. H., Harmeling, S., Macke, J. H., & Wichmann, F. A. (2016). Painfree and accurate bayesian estimation of psychometric functions for (potentially) overdispersed data. *Vision Research*, *122*, 105–123.

Seger, E., Ovadya, A., Garfinkel, B., Siddarth, D., & Dafoe, A. (2023). Democratising AI: Multiple meanings, goals, and methods.

Sering, K., Weitz, M., Shafaei-Bajestan, E., & Künstle, D.-E. (2022). Pyndl: Naïve discriminative learning in python. *Journal of Open Source Software*, *7*(80), 4515.

Settles, B. (2009). *Active learning literature survey*. University of Wisconsin-Madison Department of Computer Sciences.

Sheedy, J. E., Campbell, C., King-Smith, E., & Hayes, J. R. (2005). Progressive powered lenses: The minkwitz theorem. *Optometry and Vision science*, *82*(10), 916–922.

Shepard, R. N. (1962). The analysis of proximities: Multidimensional scaling with an unknown distance function. i. *Psychometrika*, *27*(2), 125–140.

Shepard, R. N. (1964). Attention and the metric structure of the stimulus space. *Journal of Mathematical Psychology*, *1*(1), 54–87.

Shepard, R. N. (1965). Approximation to uniform gradients of generalization by monotone transformations of scale. *Stimulus generalization*, 94–110.

Shepard, R. N. (1981). Psychological relations and psychophysical scales: On the status of "direct" psychophysical measurement. *Journal of Mathematical Psychology*, *24*(1), 21–57.

Shepard, R. N. (1987). Toward a universal law of generalization for psychological science. *Science*, *237*(4820), 1317–1323.

Shepard, R. N., Kilpatric, D. W., & Cunningham, J. P. (1975). The internal representation of numbers. *Cognitive Psychology*, *7*(1), 82–138.

Shi, Y., Bellet, A., & Sha, F. (2014). Sparse compositional metric learning. *Proceedings of the AAAI conference on artificial intelligence*, *28*.

Sievert, S., author. (2021). *Accelerating active machine learning* [Doctoral dissertation, University of Wisconsin-Madison].

Sievert, S., Nowak, R., & Rogers, T. (2023). Efficiently learning relative similarity embeddings with crowdsourcing. *Journal of Open Source Software*, *8*(84), 4517.

Sitole, S. P., LaPre, A. K., & Sup, F. C. (2020). Application and evaluation of lighthouse technology for precision motion capture. *IEEE Sensors Journal*, *20*(15), 8576–8585.

Sobol, M. (1993). Sensitivity estimates for nonlinear mathematical models. *Mathematical Modelling Computational Experiments*, *1*(4), 407–414.

Spence, I., & Graef, J. (1974). The determination of the underlying dimensionality of an empirically obtained matrix of proximities. *Multivariate Behavioral Research*, *9*(3), 331–341.

Stefan, A. M., & Schönbrodt, F. D. (2023). Big little lies: A compendium and simulation of p-hacking strategies. *Royal Society Open Science*, *10*(2), 220346.

Stevens, S. S. (1957). On the psychophysical law. *Psychological Review*, *64*(3), 153–181.

Stevens, S. S. (1959). Cross-modality validation of subjective scales for loudness, vibration, and electric shock. *Journal of experimental psychology*, *57*(4), 201–209.

Stevens, S. S. (1960). The psychophysics of sensory function. *American Scientist*, *48*(2), 226–253.

Stevens, S. S. (1946). On the theory of scales of measurement. *Science (New York, N.Y.)*, *103*(2684), 677–680.

Sullivan, C. M., & Fowler, C. W. (1988). Progressive addition and veriables focus lenses: A review. *Ophthalmic and Physiological Optics*, *8*(4), 402–414.

Sundberg, L., & Holmström, J. (2023). Democratizing artificial intelligence: How no-code AI can leverage machine learning operations. *Business Horizons*, *66*(6), 777–788.

Suppes, P., & Krantz, D. H. (2007). *Foundations of measurement: Geometrical, threshold, and probabilistic representations* (Vol. 2). Courier Corporation.

Suppes, P., & Zinnes, J. L. (1963). Basic meaurement theory. In *Handbook of mathematical psychology*. Wiley.

Suresh, S., Mukherjee, K., Padua, L., & Rogers, T. T. (2023). Behavioral estimates of conceptual structure are robust across tasks in humans but not large language models. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *45*(45).

Tabaghi, P., Peng, J., Milenkovic, O., & Dokmanić, I. (2021). Geometry of similarity comparisons. *arXiv:2006.09858 [cs, eess, stat]*.

Takane, Y. (1978). A maximum likelihood method for nonmetric multidimensional scaling. *Japanese Psychological Research*, *20*(1), 7–17.

Tamuz, O., Liu, C., Belongie, S., Shamir, O., & Kalai, A. T. (2011). Adaptively learning the crowd kernel. *Proceedings of the 28th International Conference on International Conference on Machine Learning*, 673–680.

Tanner Jr., W. P., & Swets, J. A. (1954). A decision-making theory of visual detection. *Psychological Review*, *61*(6), 401–409.

Terada, Y., & von Luxburg, U. (2014). Local ordinal embedding. *International Conference on Machine Learning*, 847–855.

Tharp, T. (2006). *The creative habit: Learn it and use it for life*. Simon & Schuster.

Thurstone, L. L. (1927). A law of comparative judgment. *Psychological Review*, *34*(4), 273–286.

Tibshirani, B. E., R. J. (1994, May 15). *An introduction to the bootstrap*. Chapman; Hall/CRC.

Torgerson, W. S. (1952). Multidimensional scaling: I. theory and method. *Psychometrika*, *17*(4), 401–419.

Toscani, M., Guarnera, D., Guarnera, G. C., Hardeberg, J. Y., & Gegenfurtner, K. R. (2020). Three perceptual dimensions for specular and diffuse reflection. *ACM Transactions on Applied Perception*, *17*(2), 6:1–6:26.

Tversky, A., & Gati, I. (1982). Similarity, separability, and the triangle inequality. *Psychological review*, *89*(2), 123.

Umbach, N. (2014). *Dimensionality of the perceptual space of achromatic surface colors* (1st ed.). Hut.

van der Maaten, L., & Weinberger, K. (2012). Stochastic triplet embedding. *International Workshop on Machine Learning for Signal Processing*, 1–6.

van Assen, J. J. R., & Pont, S. C. (2022). Identifying the behavioural cues of collective flow perception. *Journal of Vision*, *22*(14), 3985.

Vankadara, L. C., Haghiri, S., Lohaus, M., Wahab, F. U., & von Luxburg, U. (2021). Insights into ordinal embedding algorithms: A systematic evaluation. *arXiv:1912.01666 [cs, stat]*.

Victor, J. D., Aguilar, G., & Waraich, S. A. (2023, October 11). Ordinal characterization of similarity judgments.

Vincent, J. (2017, August). *Partial independence of brightness induction and brown inducton suggests a two-stage model for brightness induction* [Thesis].

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., … van Mulbregt, P. (2020). SciPy 1.0: Fundamental algorithms for scientific computing in python. *Nature Methods*, *17*(3), 261–272.

Wagenmakers, E.-J. (2007). A practical solution to the pervasive problems ofp values. *Psychonomic Bulletin & Review*, *14*(5), 779–804.

Waraich, S. A., & Victor, J. D. (2022). A Psychophysics Paradigm for the Collection and Analysis of Similarity Judgments. *JoVE (Journal of Visualized Experiments)*, (181), e63461.

Waraich, S. A., & Victor, J. D. (2024). The geometry of low- and high-level perceptual spaces. *Journal of Neuroscience*, *44*(4).

Watson, A. (1993). *Digital images and human vision*. MIT Press.

Watson, A. B., & Pelli, D. G. (1983). Quest: A bayesian adaptive psychometric method. *Perception & Psychophysics*, *33*(2), 113–120.

Weinberg, S. L., Carroll, J. D., & Cohen, H. S. (1984). Confidence regions for INDSCAL using the jackknife and bootstrap techniques. *Psychometrika*, *49*(4), 475–491.

Weinberger, K. Q., Blitzer, J., & Saul, L. (2005). Distance metric learning for large margin nearest neighbor classification. *Advances in Neural Information Processing Systems*, *18*.

Westfall, H. A., & Lee, M. D. (2021). A model-based analysis of the impairment of semantic memory. *Psychonomic Bulletin & Review*, *28*(5), 1484–1494.

Wichmann, F. A., & Hill, N. J. (2001a). The psychometric function: I. fitting, sampling, and goodness of fit. *Perception & Psychophysics*, *63*(8), 1293–1313.

Wichmann, F. A., & Hill, N. J. (2001b). The psychometric function: II. bootstrap-based confidence intervals and sampling. *Perception & Psychophysics*, *63*(8), 1314–1329.

Wichmann, F. A., & Jäkel, F. (2018, March). Methods in psychophysics. In J. T. Wixted (Ed.), *Stevens' handbook of experimental psychology and cognitive neuroscience* (pp. 1–42). John Wiley & Sons, Inc.

Wichmann, F. A., Janssen, D. H., Geirhos, R., Aguilar, G., Schütt, H. H., Maertens, M., & Bethge, M. (2017). Methods and measurements to compare men against machines. *Electronic Imaging*, *2017*(14), 36–45.

Wijntjes, M. W. A., Spoiala, C., & Ridder, H. d. (2020). Thurstonian scaling and the perception of painterly translucency. *Art & Perception*, *8*(3), 363–386.

Wills, J., Agarwal, S., Kriegman, D., & Belongie, S. (2009). Toward a perceptual space for gloss. *ACM Transactions on Graphics*, *28*(4), 1–15.

Zamir, S. W., Vazquez-Corral, J., & Bertalmio, M. (2021). Vision models for wide color gamut imaging in cinema. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *43*(5), 1777–1790.

Zhao, Y., de Ridder, H., Stumpel, J., & Wijntjes, M. (2023). Perceiving style at different levels of information. *Journal of Vision*, *23*(9), 5388.

Zhou, Y., Smith, B. H., & Sharpee, T. O. (2018). Hyperbolic geometry of the olfactory space. *Science Advances*, *4*(8), eaaq1458.

Zinnes, J. L., & MacKay, D. B. (1983). Probabilistic multidimensional scaling: Complete and incomplete data. *Psychometrika*, *48*(1), 27–48.