Similarity Ratings Reveal Expert-Novice Differences in Knowledge Organization

Priya B. Kalra[1], Heather L. Petcovic, & John Paul Minda
[1]Corresponding author: pkalra7@uwo.ca

Abstract

In this study, we examine expert-novice differences in knowledge organization using similarity ratings. Geology novices and experts provided pairwise similarity ratings for images of typical and atypical geological specimens drawn from the Rocks-30 and Rocks-360 stimulus sets provided by Nosofsky and colleagues (2018a). The pairwise similarity ratings for each group were used to construct multidimensional space representations as well as network models. Both network and MDS analyses revealed clear taxonomic differentiation by the experts. The novice MDS dimensions were easily matched to perceptual feature ratings provided in the Rocks-30 dataset. However, some of the expert dimensions did not match either the perceptual feature ratings or expert self-reported features. These results reveal expert-novice differences in knowledge organization using a quantitative methodology, and also suggest that aspects of expert knowledge organization may be missed when using methods that depend on self-report.

Keywords: expertise; knowledge organization; tacit knowledge; similarity; multidimensional scaling, network analysis

Background

Several now-classic studies on the nature of expertise and expert-novice differences established the following now-widely accepted tenets: a) that expert knowledge is structured differently from novice knowledge of the domain (i.e. hierarchically and taxonomically) b) that experts selectively attend to relevant (diagnostic) features and ignore irrelevant features and c) related to both a and b above, experts perceive stimuli from their domain of expertise differently compared to novices (see Chi, 2011 and Feltovich, 2006 for reviews). In addition to expanding our fundamental understanding of human cognition, it was thought that these breakthroughs would lead to methods for accelerating the development of expertise—if expert knowledge could be mapped, then perhaps it could be taught directly to novices, circumventing the usually required thousands of hours of experience. Sadly, these attempts at new instructional methods to accelerate expertise were not successful (e.g., Klein & Hoffman, 1993), shedding doubt on the proposition that expertise can be accelerated at all.

However, one possible reason for these failures may stem from methodological limitations of the time when the original studies were conducted. Other classic accounts of expertise emphasize that much of expert knowledge may be tacit, implicit, or automatized (Reber, 1996; Sternberg & Horvath, 1999)—in other words, not verbalizable, and often not available to expert awareness. At the same time, the methods used to map expert knowledge relied heavily on expert self-reports and/or solicitation of expert knowledge with methods designed to capture only explicit/verbalizable knowledge. This can be seen in the many discipline-based education studies using Pathfinder or similar methods to map expert knowledge or compare expert and novice knowledge structure (Neiles et al., 2016; Schvaneveldt et al., 1985; Gulacar et al., 2022)

Modern computational cognitive science approaches that use similarity as a proxy for psychological distance (Shepard, 1962, 1980) and even precedents in studying expertise (Shepard et al., 1975; Schvaneveldt et al., 1985) may provide methods for soliciting tacit expert knowledge. Importantly, not all similarity-based approaches are well suited for soliciting implicit as well as explicit aspects of knowledge. For example, because the spatial arrangement method (SpAM, Hout et al. 2013; Goldstone, 1994) asks participants to directly arrange stimuli according to similarity, it may systematically underestimate the number of underlying dimensions (Verheynen et al., 2016), and given the explicit nature of the task, this may be especially true for implicit or tacit dimensions. Thus, it is not clear if studies of expertise that use SpAM (Wing et al., 2022) capture implicit aspects of expertise.

Modern computational cognitive science methods were used by Nosofsky and colleagues (2018a) to develop a rich dataset based on the perceptual features of geological specimen images, i.e., rock pictures. Specifically, the dataset includes 360 rock pictures, evenly spanning 10 subtypes each of the 3 major rock types (Igneous, Metamorphic, and Sedimentary), feature ratings for 19 pre-determined features, pairwise similarity ratings for a set of 30 (1 per subtype) rock pictures (Rocks-30), pairwise similarity ratings for the full set of 360 images (Rocks-360), and multidimensional scaling solutions (MDS, see Shepard, 1962) based on the pairwise ratings. In the original study, the authors were able to identify most of the MDS-derived dimensions with the pre-determined features using the feature ratings. Learning to identify and classify different types of rocks is a key component of introductory geology courses, and the authors used this dataset as the basis for several instructional intervention studies (e.g., Nosofsky et al., 2018b). However, to date, this group has only applied their methods to geology novices.

In the current study, we adapt the Rocks-30 stimulus set and methodology to compare the structure of expert and novice knowledge. In doing so, we have three goals: First, to

In G. Boleda, I. Dautriche, M. Maldonado, C. Saldana and T. Torrent (Eds.), *Proceedings of the 48th Annual Conference of the Cognitive Science Society*

investigate whether modern computational methods confirm the canonical findings of expert-novice differences in knowledge structure. We predict that novice participants will prioritize visible surface features in their similarity ratings whereas expert participants will incorporate non-perceptual features, including taxonomic information, in their ratings. Second, to probe whether these methods reveal aspects of expert knowledge structures not available from self-report-based measures (i.e., implicit aspects of expert knowledge). Finally, in keeping with the original goals of the Rocks-30 and Rocks-360 study, we aim to create a basis for future discipline-based research in geology education using what we learn about expert knowledge this way.

Methods

Procedure

Two groups of participants were recruited: geology experts and novices. Novices completed either the typical exemplar rating task or the atypical exemplar rating task. Experts completed the typical exemplar rating task and a (self selected) subset of those experts also completed the atypical exemplar rating task and/or the feature generation task.

Stimuli

All stimuli were taken from the Rocks 360 stimulus set provided by Nosofsky et al. on the OSF platform. This rich stimulus set includes not only visual stimuli (i.e., pictures of rocks: 12 exemplars for each of 30 rock subtypes) but also perceptual feature ratings for each stimulus. Stimuli for the typical exemplar ratings were taken from the Rocks-30 subset of these stimuli identified by Nosofsky et al. Atypical exemplars were selected from among the remaining exemplars in each rock subtype as follows: first, the perceptual feature ratings were used to calculate L2 norms (i.e. perceptual distance across all perceptual feature dimensions) for each pair of exemplars within a category. Next, these pairwise distances were averaged for each exemplar. The exemplar with the highest mean L2 norm (i.e., highest mean perceptual distance from the other exemplars) was selected as the atypical exemplar for that rock subtype.

Participants

Novices were recruited from a participant pool of undergraduate students completing required psychology research participation and were compensated with participation credits. Lack of geology expertise was established by self-report to a questionnaire about students' experience with/knowledge of geology. (M_age = 18.5 years, 59% Female, 39% Male, 2% Self-identify/nonbinary) 212 participants completed all trials of the typical exemplar rating task; a smaller sample of 114 participants completed all trials of the atypical exemplar rating task.

Experts were recruited by sending an email with a link to the experiment to several email lists used by geology professionals (Geological Society of America, the American Geophysical Union, and the National Association of Geoscience Teachers). Expertise was confirmed by 1) a rock identification task and 2) a questionnaire on expert participants' knowledge of geology, occupational sector, and experience with rock identification. Experts were compensated with a $5 Amazon gift card for each task completed as well as a raffle entry for each task completed. Age and gender information were not collected for experts.

Tasks

Participants received a link, either through SONA or through email, to a Qualtrics form. The Qualtrics included consent, embedded tasks, questionnaires, and debriefing information. The embedded tasks were created using PsychoPy and uploaded to Pavlovia for online use and distribution.

Typical and atypical exemplars similarity rating. Following Nosofsky et al. (2018), on each trial, participants saw two rock pictures and a scale with anchors at 1 (left) and 9 (right) and were asked to rate the similarity of the two pictures where 1=not similar and 9=very similar (with no time limit per trial). No definition of similarity was given. Visual white noise masks were presented in place of each of

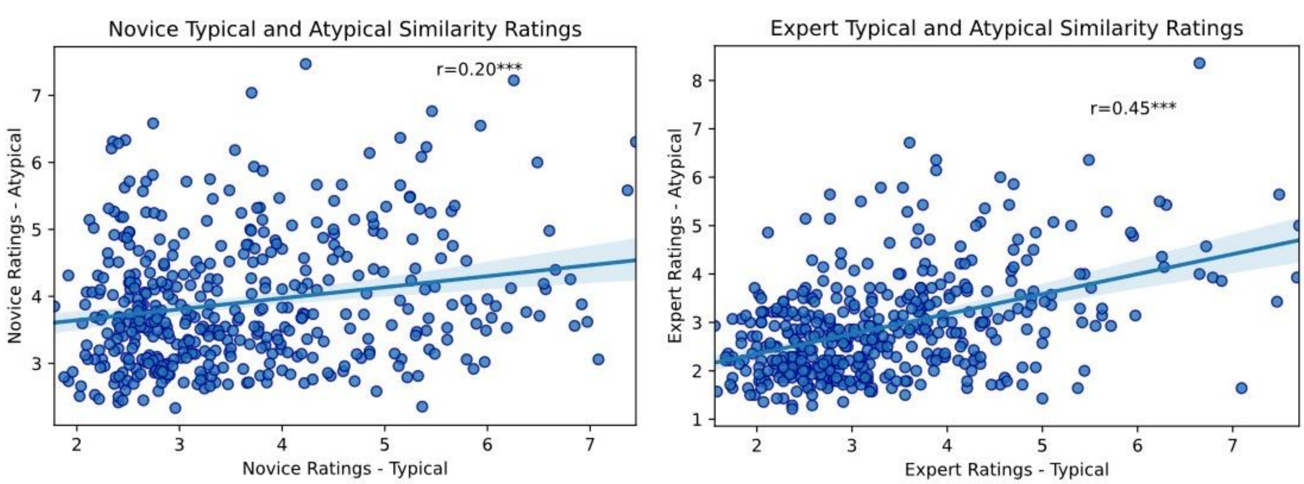

Figure 1. Scatterplots of typical exemplar and atypical exemplar pairwise similarity ratings for novices and experts.

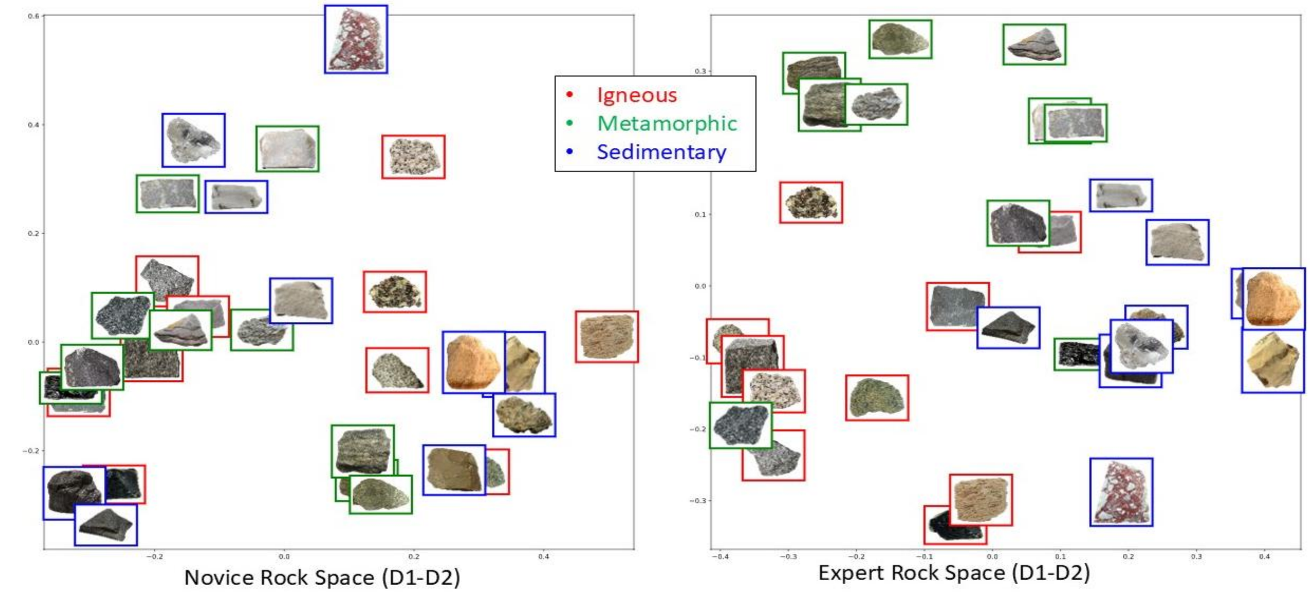

Figure 2. Dimensions 1 and 2 of MDS solution for novices (left) and experts (right)

the pictures for 200 msec between trials. All participants completed 435 similarity rating trials (every unique pair of the 30 images: $(n^2-n)/2$) broken into 15 blocks, with breaks available between blocks. Order of blocks and order of trials within block were randomized.

Feature generation task (experts only) After they completed the similarity rating tasks, experts were prompted to provide free typed responses to the question "What features or characteristics do you use to distinguish between different types of rocks? (please list as many as you would like, separated by commas)."

Analysis

Multidimensional Scaling For each group, mean pairwise similarity ratings were computed. The mean pairwise similarity ratings for typical and atypical exemplars of the same type of rock were correlated separately for novices and experts. Again following Nosofsky et al., (2018), non-metric multidimension scaling was applied to the mean pairwise similarity ratings and the resulting solutions were subsequently rotated using principal components analysis to obtain a more interpretable coordinate system. Unlike Nosofsky et al., (2018), no Procrustes transformations were applied.

Network analysis We constructed an undirected network from the pairwise similarity matrix for each group; each node represents a typical exemplar rock used in the rating task. Each group's similarity matrix was thresholded using the mean similarity value for that group; only values greater than the mean were retained. The remaining mean pairwise similarity ratings were used as edge weights. Community structure was assessed using modularity optimization, specifically using the "modularity" function in the *networkx* Python package (Hagberg et al., 2008).

Results

Novice replication

Mean pairwise similarity ratings from our novice-typical exemplar sample correlated with the original Nosofsky et al. (2018) pairwise similarity at r=.89, p<.001, allowing us to infer that we accurately replicated the methods of the original study, and providing support for the replicability of the original study.

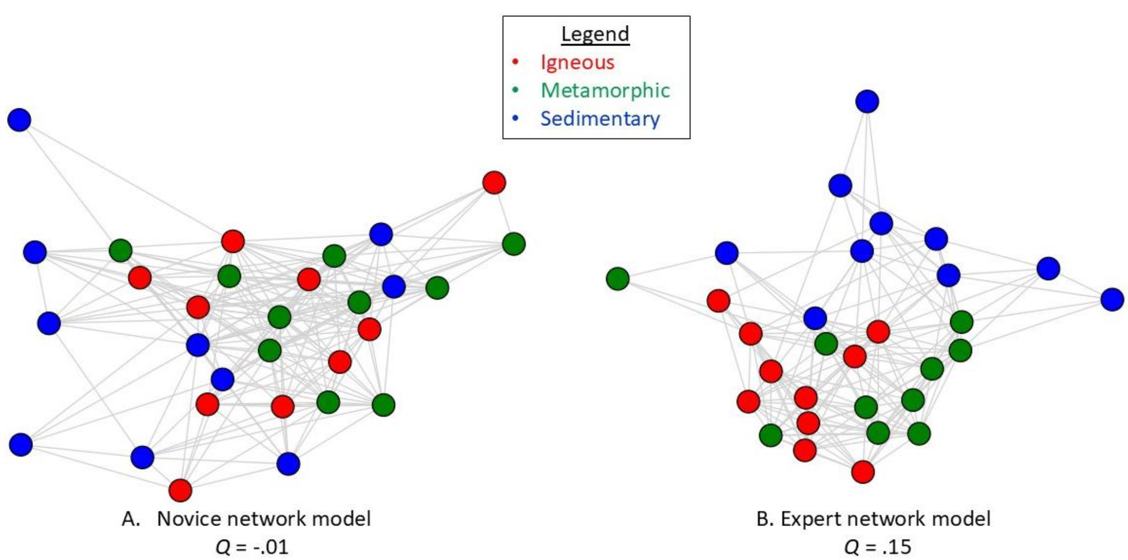

Figure 3. Network models based on novice and expert similarity ratings. Q = modularity based on rock type partition.

Typical-Atypical Correlations

Pairwise similarity ratings for typical exemplar pairs and atypical exemplar pairs were correlated within each group (novices and experts) (Figure 1). The correlation between typical and atypical exemplar ratings for novices (r = .20, p<.001) was significantly less than the same correlation for experts (r = .45, p<.001) as determined by applying Fisher's z transformation method (z=4.14, p<.001).

Multidimensional Scaling

Following the procedure in the original study, we applied multidimensional scaling to the pairwise similarity data (typical exemplars only). For the novices, a model with 8 dimensions was found to have the best fit; for the experts a 12-dimensional model was found to have the best fit based on stress (see Appendix). We then plotted the dimensions in pairs (e.g., dimensions 1 and 2, dimensions 3 and 4). As seen in Figure 2, clear separation between rock types (Igneous, Metamorphic, Sedimentary) can be seen in the plot of expert MDS dimensions 1 and 2; no such separation was found in any pair of novice dimensions (dimensions 1 and 2 are displayed for comparison), nor in a 3d plot of the first 3 novice MDS dimensions.

Modularity

Figure 3 displays the novice and expert networks. A modularity quotient (Q) was computed for each network, using the three rock types to partition the network into communities. For the novices, Q = -.01, suggesting little to no community structure. For experts, Q =.15, indicating a greater degree of community structure based on rock types, which is confirmed by visual inspection of the network models.

Expert feature generation

Expert responses were coded and tabulated by frequency, with a weight measure computed as the inverse of order of production (i.e., the first response produced would have the greatest weight). The 10 most frequent coded responses are shown in Table 1.

Table 1: Expert-generated features

Feature Code	Frequency	Weight
Grain size	32	495
Texture	29	519
Minerals/Mineralogy	28	471
Color	26	434
Layering	23	321
Lithology	14	226
Composition	13	226
Taste, Feel, Smell	12	158
Crystals/glass	11	169
Grain shape	11	159

Interpreting MDS Dimensions

Each of the MDS dimensions was compared to the feature ratings provided from the original study for experts and novices separately.

Novices Correlations between novice MDS dimensions and feature ratings are displayed in Table 2. Our results generally replicate the pattern found in the original study. We confirmed that the correlation values reflected a linear trend for the first 6 dimensions rather than the influence of outliers by visual inspection of feature by dimension scatterplots (see online supplemental materials at https://osf.io/zfmbe/overview?view_only=a4306bff12d549b2acd2d0c1243ab469). We note that the correlation values may be lower because we did not perform the Procrustes transformations.

Table 2: Maximum correlation between each novice MDS dimension and rated features

MDS dimension	Feature	r	p
dim1	saturation	0.80	0.00
dim2	lightness	0.61	0.00
dim3	grain size	-0.71	0.00
dim4	shininess	0.74	0.00
dim5	color (rho)[1]	0.46	0.01
dim6	organized	-0.48	0.01
dim7	holes	-0.31	0.09
dim8	special	-0.39	0.04

Experts Correlations between expert MDS dimensions and feature ratings are displayed in Table 3. Linear trends determined by visual inspection were found for dimensions 1 (grain size), 3 (lightness of color) , 4 (shininess), 5 (organized), 8 (layers) and 9 (layers) (see online supplemental materials for scatterplots). We plotted dimensions 2, 6, and 7 against each of the provided feature ratings in the original study, but found no linear trends or significant correlations. The MDS derived dimensions in Table 3 can be compared to the expert-generated features in Table 1.

Table 3: Maximum correlation between each expert MDS dimension and rated features

MDS dimension	Feature	r	p
1	grain size	-0.49	0.01
2	stripes	0.43	0.02
3	lightness	-0.72	0.00
4	shininess	0.57	0.00
5	organized	-0.68	0.00
6	holes	0.53	0.00
7	stripes	-0.38	0.04
8	layers	0.38	0.04
9	layers	-0.46	0.01
10	color variation	-0.31	0.09

[1] Briefly, we transformed the provided cartesian (x-y) coordinates (where distance from origin ~ chroma and angle ~ hue) to polar coordinates, so that rho indicates distance from origin and phi indicates hue angle. Rho here is roughly interpretable as a transformed saturation value.

11	organized	-0.33	0.08
12	holes	0.25	0.18

Discussion

Using modern quantitative methods, we have confirmed canonical findings regarding expert knowledge organization and expert-novice differences in knowledge organization in several ways. First, the relatively high correlation for experts between pairwise similarity for perceptually different exemplar pairs belonging to the same categories reflects the notion that experts are able to ignore irrelevant perceptual features and apply knowledge of the underlying taxonomic structure of the domain. Only some of this correlation can be explained by within-category perceptual similarity, which can be seen in the novice typical-atypical pair correlations. Second, plotting the first two expert MDS dimensions reveals a space in which the exemplars are highly separable based on their high-level rock type (igneous, metamorphic, or sedimentary), which reflects experts' use of taxonomic knowledge and hierarchical organization. Third, we quantified this separation as modularity scores in a network model, again showing organization based on rock type in experts (but not novices).

Importantly, expert generated features and the expert MDS dimensions differ in the importance they assign to different features. Although grain size is first in both lists, they diverge after that. Notably, shininess (or lustre) was reported by only a small number of experts in self-report, but through MDS shininess was demonstrated to explain significant variance in expert similarity ratings. Also notable, MDS expert dimension 2 could not be correlated with any of the provided perceptual feature ratings. While these discrepancies may be due to measurement or modeling issues, it is also possible that the similarity ratings-MDS approach provides information that self-report alone could not, such as tacit or implicit expert knowledge.

Interestingly, many of the expert generated features are not easily perceived in pictures of rocks (such as mineral composition or lithology), but reflect underlying taxonomic knowledge and knowledge of rock formation processes. However, some of these features may potentially be inferred from perceptual features: for example, grain size is highly predictive of lithology, and color may be informative regarding mineral composition. In some of the feature generation responses, as well as informal debriefing, experts describe a two-step process in which they identified or classified the rock pictures before determining similarity, as well as conditional use of certain features (for example, if the rock was identified as igneous, only then was lightness used). Future studies and analyses may be able to model the complexity of expert processing with more nuance, for example using vector-symbolic algebra.

The results of the current study provide a foundation for future observational and intervention studies of rock identification learning. Now that the expert and novice rock spaces have been mapped, a clear next step is to observe the development of expert knowledge organization, for example by comparing rock spaces from a cross-section of geology students with different degrees of expertise. The results of such a study could reveal whether the progression from novice to expert knowledge organization is gradual and continuous or whether there are abrupt qualitative reorganizations.

The mapping of expert rock space also provides domain-specific information that could be used to design paradigms that might accelerate the development of rock identification expertise. For example, the finding that grain size is the top feature used by experts in both self-report and similarity ratings suggests that directing students to attend to this feature and place less weight on color-related features (i.e., feature highlighting) could improve novice rock identification and classification[2]. Another potential approach could involve the use of conditional or step-wise application of perceptual features. Finally, linking the perceptual features (e.g. grain size) with taxonomic domain knowledge (rock creation is causally related to both type—Igneous, Metamorphic, or Sedimentary—and grain size) could prove to be another route, especially if perceptual and conceptual learning can be linked as recent studies suggest (Kalra et al., 2025).

In summary, quantitative methods that are not heavily dependent on awareness may capture aspects of expert knowledge that other methods (such as Pathfinder) may not. Depending on the size of the stimulus set, and given optimization methods for gathering similarity ratings (e.g. Sievert et al., 2023), data collection need not be burdensome. By using a map of expert knowledge that includes potentially tacit knowledge as a goal state, theoretically motivated approaches to improving instruction, whether based in human category learning (Nosofsky et al., 2018b,c) or machine teaching (Zhu, 2015) can quantify the effectiveness of different instructional strategies in moving novices toward expert-like knowledge organization.

Acknowledgments

The authors declare that no AI was used in the preparation of this manuscript.

[2] Although attempts have been made to apply feature highlighting to rock identification training with varying degrees of success (Meagher et al., 2022; Miyatsu et al., 2019), we note that the technique used is these studies is not entirely aligned with other instantiations of feature highlighting (for example Mac Aodha, 2018), so additional attempts are warranted.

Appendix

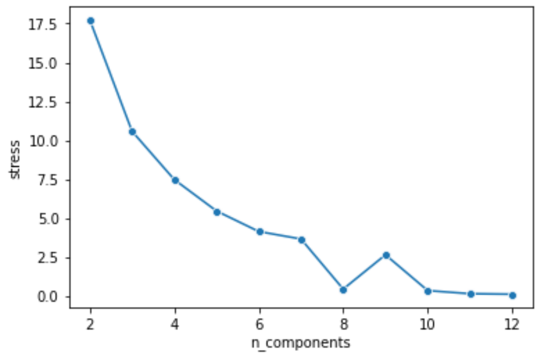

Figure A1. Novice MDS stress plot.

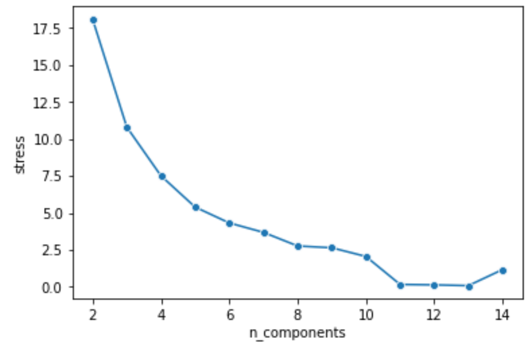

Figure A2. Expert MDS stress plot.

MDS dimensionality selection. For both experts and novices, the model selected for further analysis was the model that minimized stress and/or indicated an asymptote in stress reduction. Figures A1 and A2 display plots of stress against model (number of components).

References

Chi, M. T. H., Feltovich, P. J., & Glaser, R. (1981). Categorization and Representation of Physics Problems by Experts and Novices. *Cognitive Science*, *5*(2), 121–152. https://doi.org/10.1207/s15516709cog0502_2

Chi, M. T. H. (2011). Theoretical perspectives, methodological approaches, and trends in the study of expertise. In Expertise in Mathematics Instruction: An International Perspective (pp. 17–39). Springer US. https://doi.org/10.1007/978-1-4419-7707-6_2

Feltovich, P. J., Prietula, M. J., & Ericsson, K. A. (2012). Studies of Expertise from Psychological Perspectives. In *The Cambridge Handbook of Expertise and Expert Performance* (pp. 41–68). https://doi.org/10.1017/cbo9780511816796.004

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

Gulacar, O., Vernoy, B…Milkey, A. (2022) Investigating differences in experts' chemistry knowledge structures and comparing them to those of general chemistry students. *Journal of Chemical Education 99* (8), 2950-2963 DOI: 10.1021/acs.jchemed.2c00251

Hagberg,A., Schult. D.A.,& Swart, P.J. (2008) "Exploring network structure, dynamics, and function using NetworkX", in Proceedings of the 7th Python in Science Conference (SciPy2008), Gäel Varoquaux, Travis Vaught, and Jarrod Millman (Eds), (Pasadena, CA USA)

Hout, M. C., Goldinger, S. D., & Ferguson, R. W. (2013). The versatility of SpAM: A fast, efficient, spatial method of data collection for multidimensional scaling. *Journal of Experimental Psychology: General*, 142(1), 256.

Kalra, P. B., Batterink, L. J., & Minda, J. P. (2025). Procedural and declarative knowledge simultaneously contribute to category response selection. *Psychological Research*, *89*(5), 146.

Klein, G. A., & Hoffman, R. R. (1993). Seeing the invisible: Perceptual-cognitive aspects of expertise. In M. Rabinowitz, (Ed.), *Cognitive science foundations of instruction*. Hillsdale, NJ: Lawrence Erlbaum Assoc

Mac Aodha, O., et al. (2018). Teaching categories to human learners with visual explanations. *CVPR*

Mazzuca, C., & Majid, A. (2023). The semantic representation of food is shaped by cultural experience. *Language and Cognition*, *15*(4), 651-669.

Miyatsu, T., Gouravajhala, R., Nosofsky, R. M., & McDaniel, M. A. (2019). Feature highlighting enhances learning of a complex natural-science category. *Journal of Experimental Psychology: Learning Memory and Cognition*, *45*(1), 1–16. https://doi.org/10.1037/xlm0000538

Neiles, K. Y., Todd, I., & Bunce, D. M. (2016). Establishing the Validity of Using Network Analysis Software for Measuring Students' Mental Storage of Chemistry Concepts. Journal of Chemical Education, 93(5), 821-831.

Meagher, B. J., McDaniel, M. A., & Nosofsky, R. M. (2022). Effects of Feature Highlighting and Causal Explanations on Category Learning in a Natural-Science Domain. Journal of Experimental Psychology: Applied, 28(2), 283–313. https://doi.org/10.1037/xap0000369

Norman, G. R., & Brooks, L. R. (1997). The Non-Analytical Basis of Clinical Reasoning. In *Advances in Health Sciences Education* (Vol. 2). Kluwer Academic Publishers.

Nosofsky, R. M., Sanders, C. A., & McDaniel, M. A. (2018b). A Formal Psychological Model of Classification Applied to Natural-Science Category Learning. Current Directions in Psychological Science, 27(2), 129–135. https://doi.org/10.1177/0963721417740954

Nosofsky, R. M., Sanders, C. A., Meagher, B. J., & Douglas, B. J. (2018a). Toward the development of a feature-space representation for a complex natural category domain. Behavior Research Methods, 50(2), 530–556. https://doi.org/10.3758/s13428-017-0884-8

Nosofsky, R. M., Sanders, C. A., Zhu, X., & McDaniel, M. A. (2018c). Model-guided search for optimal natural-science-category training exemplars: A work in progress. Psychonomic Bulletin & Review. https://doi.org/10.3758/s13423-018-1508-8

Reber, A. S. (1989). Implicit Learning and Tacit Knowledge. *Journal of Experimental Psychology-General*, 118(3), 219–235. https://doi.org/10.1037//0096-3445.118.3.219

Schvaneveldt, R. W., Durso, F. T., Goldsmith, T. E., Breen, T. J., Cooke, N. M., Tucker, R. G., & De Maio, J. C. (1985). Measuring the structure of expertise. *International journal of man-machine studies*, *23*(6), 699-728.

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

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

Shepard, R. N. (1980). Multidimensional scaling, tree-fitting, and clustering. *Science*, 210(4468), 390–398.

Sievert et al., (2023). Efficiently Learning Relative Similarity Embeddings with Crowdsourcing. *Journal of Open Source Software*, 8(84), 4517, https://doi.org/10.21105/joss.04517

Sternberg, R. J., & Horvath, J. A. (Eds.). (1999). *Tacit knowledge in professional practice: Researcher and practitioner perspectives*. Psychology Press.

Verheyen, S., Voorspoels, W., Vanpaemel, W., & Storms, G. (2016). Caveats for the spatial arrangement method: Comment on Hout,Goldinger, and Ferguson (2013). *Journal of Experimental Psychology: General*, 145, 376–382.

Wing, E. A., Burles, F., Ryan, J. D., & Gilboa, A. (2022). The structure of prior knowledge enhances memory in experts by reducing interference. *Proceedings of the National Academy of Sciences*, *119*(26), e2204172119.

Zhu, X. (2015). Machine teaching: An inverse problem to machine learning and an approach toward optimal education. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence* (pp. 4083–4087). AAAI Press