Cognition 238 (2023) 105511

Contents lists available at ScienceDirect

Cognition

journal homepage: www.elsevier.com/locate/cognit

Original articles

Teaching categories via examples and explanations

Arseny Moskvichev [a,c,*], Roman Tikhonov [b], Mark Steyvers [a]

[a] Department of Cognitive Sciences, University of California, Irvine, CA, United States of America
[b] Department of Social and Decision Sciences, Carnegie Mellon University, PA, United States of America
[c] Santa Fe Institute, Santa Fe, NM, United States of America

ARTICLE INFO

Dataset link: https://osf.io/hjyu5/

Keywords:
Category learning
Pedagogical communication
Category communication
Exemplar-based instruction
Verbal instruction

ABSTRACT

People often learn categories through interaction with knowledgeable others who may use verbal explanations, visual exemplars, or both, to share their knowledge. Verbal and nonverbal means of pedagogical communication are commonly used in conjunction, but their respective roles are not fully understood. In this work, we studied how well these modes of communication work with different category structures. We conducted two experiments to investigate the effect of perceptual confusability and stimulus dimensionality on the effectiveness of verbal, exemplar-based, and mixed communication. One group of participants – teachers – learned a categorization rule and prepared learning materials for the students. Students studied the materials prepared for them and then demonstrated their knowledge on test stimuli. All communication modes were generally successful, but not equivalent, with mixed communication consistently showing best results. When teachers were free to generate as many visual exemplars or words as they wish, verbal and exemplar-based communication showed similar performance, although the verbal channel was slightly less reliable in situations requiring high perceptual precision. At the same time, verbal communication was better suited to handling high-dimensional stimuli when communication volume was restricted. We believe that our work serves as an important step towards studying language as a means for pedagogical category leaning.

1. Introduction

One of the most striking features of the human mind is our ability to share knowledge with each other. Learning from direct experience takes time, effort, and might even be dangerous; learning through communication is safer and more efficient, which provides numerous benefits for humans as individuals and as a species (Bandura, 1977; Tomasello, 1999; Vygotsky, 1978). From personal experience, we know that knowledge communication can naturally be mediated through different (verbal and nonverbal) channels. For example, imagine a family forest trip with parents teaching their children about poisonous and edible mushrooms. It is easy to envision a parent instructing through verbal explanations (e.g., not to collect pale, thin-legged mushrooms with a flat cap since they are usually poisonous), thus providing a definition of a concept that can be later reused. Another way to teach the same concept is to give labeled examples: one may sort, together with the child, through the mushrooms that the child collected, keeping the good ones, and throwing away the bad ones. The key difference is that the former involves a verbal explanation, while the latter relies mostly on nonverbal (exemplar-based) pedagogical communication.

Similarly to the example above, in this paper, we focus on knowledge communication in a category learning setting. Our ability to determine category membership based on past experience is a fundamental skill, involved in many aspects of human cognitive organization, and used ubiquitously in a wide range of situations (see Ashby & Maddox, 2005; Ashby, Paul, & Maddox, 2011; Dubova & Goldstone, 2021; Seger & Miller, 2010). At the same time, category learning is extremely convenient methodologically: one gets a clear way to define what exactly is being learned, fully control the learning procedure, easily measure learning outcomes, and, finally, easily instantiate the process into a formal mathematical model. These features made category learning one of the most common approaches for studying knowledge acquisition. They also make category learning perfectly suitable for investigating knowledge communication, although researchers have only recently begun to explore this direction (e.g., Aodha, Su, Chen, Perona, & Yue, 2018; Chopra, Tessler, & Goodman, 2019; Moskvichev, Tikhonov, & Steyvers, 2019). A notable exception is a pioneering study by Avrahami et al. (1997) who introduced a teaching-by-examples paradigm and demonstrated that teacher-generated learning sequences result in higher students' performance than equivalent sets of stimuli presented in random order. This paradigm was further extended by Shafto, Goodman, and Griffiths (2014) who built a Bayesian computational model providing insight into the methods of formally describing pedagogical interaction in a category learning setting (discussed in more detail

* Corresponding author.
E-mail address: amoskvic@uci.edu (A. Moskvichev).

https://doi.org/10.1016/j.cognition.2023.105511
Received 13 April 2022; Received in revised form 5 June 2023; Accepted 5 June 2023
Available online 1 July 2023

later). They, however, focused solely on communication via selecting category examples and ignored language-based communication.

In this work, we investigate how people communicate perceptual categories using different communication channels (verbal, exemplar-based, or mixed). We have conducted two experiments, in which we varied perceptual confusability and stimulus dimensionality to capture fundamental differences between the verbal and exemplar channels of communication. In Experiment 1, we investigated how varying category structures affect communication efficiency of verbal, nonverbal (exemplar-based), and mixed teaching formats. Experiment 2 complemented our previous findings by limiting the number of examples and words that teachers were able to use to communicate categories. These constraints mitigated the variability in teachers' communication and allowed us to take a more nuanced look at the roles of different channels of communication.

1.1. Category learning in a pedagogical setting

Pedagogical learning, i.e., learning from someone who intentionally chooses teaching materials, is qualitatively different from learning categories by observing random data samples, and its modeling presents unique challenges. A solution to this problem was proposed by Shafto et al. (2014). Following the rational analysis framework (Anderson, 1990, 1991), the authors proposed and empirically validated a computational model for the process of exemplar-based pedagogical reasoning. The model is built upon the idea of mutual rationality assumption: rational teachers choose materials that would maximize a rational learners' ability to infer the categorization rule and achieve good performance. Rational learners, in turn, base their inferences by assuming that teachers are behaving rationally and are being helpful. Earlier, Avrahami et al. (1997) illustrated this "mutual rationality" idea in a series of experiments using a "teaching-by-examples" paradigm. The study revealed consistent and effective patterns of pedagogical communication employed when teachers use a sequence of visual examples to communicate their category knowledge. Similar findings have been obtained in children, where it has been shown that their decisions on what to teach are made in a way that maximizes learners' rewards (Bridgers, Jara-Ettinger, & Gweon, 2020). Overall, formal theoretical frameworks clearly demonstrate the uniqueness of the pedagogical setting in how it may affect category learning. And yet, even though much of our communication (including category communication) is pedagogical in nature and involves both verbal and exemplar-based modes of communication, the differences between these ways of teaching have received little attention (but see Sumers, Ho, Hawkins, and Griffiths (2023), for a recent example outside of category learning).

1.2. Category learning and language

Many theories of category learning agree that both verbal and nonverbal processes are involved in categorization. There is still, however, an ongoing debate on whether the verbal-like and nonverbal processes are performed by two different (Ashby et al., 1998; Ashby & Maddox, 2005; Maddox & Ashby, 2004; Minda & Miles, 2010) or only one (Keren & Schul, 2009; Newell, Dunn, & Kalish, 2011) cognitive system. Weighing in on this long-standing debate is outside the scope of our paper. Nevertheless, there is ample evidence that regardless of whether one or two systems are involved, one of them must be able to handle and utilize verbal knowledge.

Language is more than just a communication tool (see further in Clark, 1998; Gentner, 2016; Lupyan, 2012), it can reshape and facilitate category learning in many different ways. First, it can be used as a tool for labeling dimensions: a number of recent studies demonstrated that feature nameability (ease of finding verbal labels for relevant dimensions) promotes categorization performance (Kotov & Kotova, 2018; Zettersten & Lupyan, 2018, 2020). Second, language can be helpful in directing attention to the most informative stimuli features (Sloutsky, 2010; Sloutsky et al., 2016). As a result, language can be especially useful in learning categories consisting of objects with few relevant dimensions and multiple independently varying irrelevant features — i.e., statistically sparse categories as defined by Kloos and Sloutsky (2008). Finally, language can be used to account for unobservable characteristics of objects while categorizing them and forming nested categories of different abstraction levels (Sloutsky, 2010). Even though the importance of language-related processes is largely acknowledged, there is very little research into studying the properties of language as the primary means of pedagogical category communication.

1.3. Identifying factors that may differentially affect verbal and exemplar-based communication

To the best of our knowledge, no studies of category communication directly examined the factors that affect verbal and exemplar-based pedagogical communication of categories. Because of that, when looking for potential factors that might differentially affect verbal and exemplar-based communication, we had to extrapolate from category learning studies in individual settings.

We know from previous studies that categorization performance is affected by category structure (Shepard, Hovland, & Jenkins, 1961): some category structures (e.g., defined by a unidimensional rule) are more easily learned through verbal means, while others (e.g., involving a combination of multiple features or family resemblance categories) rely on procedural memory and nonverbal processes (Ashby et al., 1998; Maddox & Ashby, 2004). Overall, we believe that the latter type is not well suited for investigating in a pedagogical setting, as these categories are extremely difficult to verbalize and transfer to another person. Therefore, in our study, we focus on the first type of categories.

Categories that follow the same rule type may vary in their difficulty, depending on the perceptual similarity/confusability of its members. Perceptual similarity/confusability is usually operationalized through within-category (Cohen, Nosofsky, & Zaki, 2001; Rips, 1989; Smith & Sloman, 1994) and between-category variability. The larger the within-category variability, the harder it is to rely on prototype information when making judgments. At the same time, it may facilitate rule abstraction since rule-based categorization strategy is the most appropriate one for these categories (Kloos & Sloutsky, 2008). Between-category similarity affects categorization performance by making it difficult to determine the boundary between categories. Categories with fuzzy boundaries (i.e., with many borderline examples of different categories located close to each other) are naturally more challenging to learn. Categorization difficulty is also related to the number of irrelevant dimensions varied within a category. Stimuli with multiple irrelevant dimensions require larger training samples or additional efforts to direct attention to the relevant dimension while ignoring the rest of the information (e.g., Vong, Hendrickson, Navarro, and Perfors, 2019).

Kloos and Sloutsky (2008) combined perceptual similarity and dimensionality metrics to calculate statistical density of categories. Statistically dense categories are the ones that have multiple relevant covarying features that determine category membership. They also have lower within-category variability and higher between-categories distinctiveness. Sparse categories, on the contrary, have multiple independently varying irrelevant dimensions and only few dimensions that determine category membership. Statistically dense categories are better learned in nonverbal manner — by mere observation of category examples, while sparse categories require prior verbal instruction to constrain learner's hypothesis space and enable selective attention to the relevant features (see also Aboody, Velez-Ginorio, Laurie, Santos, and Jara-Ettinger, 2018).

Based on these prior results, we formulated a number of hypotheses. First, we expected that the relative efficiency of teaching via verbal explanations would increase with higher stimuli dimensionality (compared to exemplar-based teaching). Second we expected that higher

confusability would increase the relative efficiency of teaching via visual exemplars (compared to verbal explanations). Lastly, we also expected emergent effects when using two channels of communication simultaneously; specifically, we hypothesized that communication of categorical information will be more efficient (per communication unit) if verbal explanations are combined with learning-by-examples (compared to verbal explanations or examples alone).

Here, it is important to mention that these main hypotheses were pre-registered before the pilot study (see Appendix B). The pilot study strongly suggested that our original hypotheses, especially the one concerning the mixed channel, were only likely to hold if re-formulated in terms of effectiveness, as opposed to efficiency (accuracy per communication unit) which was used in the pre-registration. Based on the pilot, it also became clear that accuracy per communication unit, as a metric, is very sensitive to the choice of a conversion method between words and examples. During the pilot, we, therefore, switched to separately analyzing communication effectiveness as our main quantity of interest, complementing it with a separate communication volume analysis. We used the same approach in our main experiments.

1.4. Overview of the experiments

We conducted two experiments to investigate the effects of category structure and communication channels on category communication effectiveness and efficiency in a teacher–student format. In both experiments, one group of participants (*teachers*) learned a categorization rule and prepared learning materials for another group (*students*). Students studied the materials prepared for them and then demonstrated their knowledge on test stimuli. In Experiment 1, we varied perceptual confusability (high vs. low) and stimulus dimensionality (two, three, or four dimensions) to study the differences between three communication formats (verbal explanations, visual exemplars, or a mixture of both). In Experiment 2, we limited the amount of materials that teachers were allowed to communicate to account for differences in teachers' efforts. We focused only on two communication channels (verbal and exemplar-based) and excluded a three-dimensional condition as uninformative. Communication was asynchronous in all experiments. Students received learning materials prepared by teachers in advance, and there were no other interactions between teachers and students.

2. Experiment 1

2.1. Method

2.1.1. Procedure

There were two groups of participants, *teachers* and *students*. For teachers, the main part of the experiment consisted of three stages: learning phase, test phase, and teaching phase (see Fig. 1). During the *learning phase*, teachers learned a specific category through 30 randomly sampled labeled examples. Stimuli were presented simultaneously so that participants could easily infer a categorization rule by observing examples at their own pace. Teachers were able to explore each stimulus in detail by enlarging it and had no time constraints. Examples of teachers' learning materials are provided in Appendix A.

Every block of 30 training examples was followed by a *test phase*, where teachers were tested on 30 new examples with no feedback. If they achieved categorization accuracy of 85% or above, the teacher proceeded to the *teaching phase*, otherwise, they returned to training. If a teacher failed to pass the test five times, the experiment ended without transitioning to the teaching phase. We used this strict accuracy threshold for teachers to minimize interference of teacher learning performance with communication efficiency and effectiveness, as well as the overall quality of their teaching materials. In other words, we wanted to see how knowledgeable teachers communicate their knowledge, and so we had to make sure that teachers master their category knowledge in all conditions before proceeding to teaching.

During the *teaching phase*, teachers generated learning materials for their future students in three different formats: verbal, exemplar-based, and mixed. The verbal format required teachers to formulate a written message with an explanation of how to distinguish between members of two categories. In the exemplar-based format, teachers generated labeled stimulus examples (separately for each of the categories) through an interface that allowed them to adjust stimulus characteristics using sliders for different features. In the mixed format, teachers were able to use a combination of exemplars and verbal explanations (see Appendix A for details on the interface). The order of teaching formats was randomized and teachers had no ability to get back and copy previously created materials. Teachers were instructed to make each set of instructions self-contained. That is, they knew that each of their students would receive only one of these three teaching materials.

For the students, the experiment was shorter. In the *learning phase*, they observed the materials prepared for them by their teacher. Just as with teachers, there was no time restriction on how long they took to study the materials. When ready, they proceeded to the test stage (containing 30 stimuli), where their mastery of the communicated category was measured. See the details on the student interface in Appendix A.

2.1.2. Design: independent variables

We used a 3 × 2 between-subject design. *Teachers* were assigned into one of six groups defined by the following category characteristics: stimulus dimensionality (two, three, or four varying dimensions), and perceptual confusability (low or high).

Stimulus dimensionality. We varied the number of dimensions along which stimuli may change (i.e. two-dimensional stimuli have two varying features). We had two-, three-, and four-dimensional stimulus conditions.

Perceptual confusability. Confusability was defined as a ratio of the gap between the categories to the variance within these categories (see Fig. 2). If the gap is large, compared to the within-category variation, it is easy to distinguish between instances of different categories. Moreover, it is likely that there is going to be a specific label one may use to indicate the threshold.

Communication format. In addition to the between-subject independent variables listed above, we also manipulated communication format as a within-subject variable (for teachers only). Each teacher was required to create teaching materials in three different formats: exemplars-only, verbal-only, and mixed. They were instructed that their students will see only one of these three materials.

Students. Each student was randomly assigned a teacher and learned from materials presented in one of three communication formats (verbal, exemplar-based, or mixed).

2.1.3. Design: dependent variables

Performance metrics. First and foremost, we were interested in participants' ability to communicate category knowledge in different conditions. To do so, we used student accuracy as the target variable. To make sure that accuracy differences reflected differences in communication effectiveness rather than differences in teachers' initial knowledge, we controlled for each student's respective teacher's accuracy (by including it as a covariate). When visualizing the data, we used the difference between each teacher's and their student's accuracies, to visually represent the amount of accuracy "lost in communication".

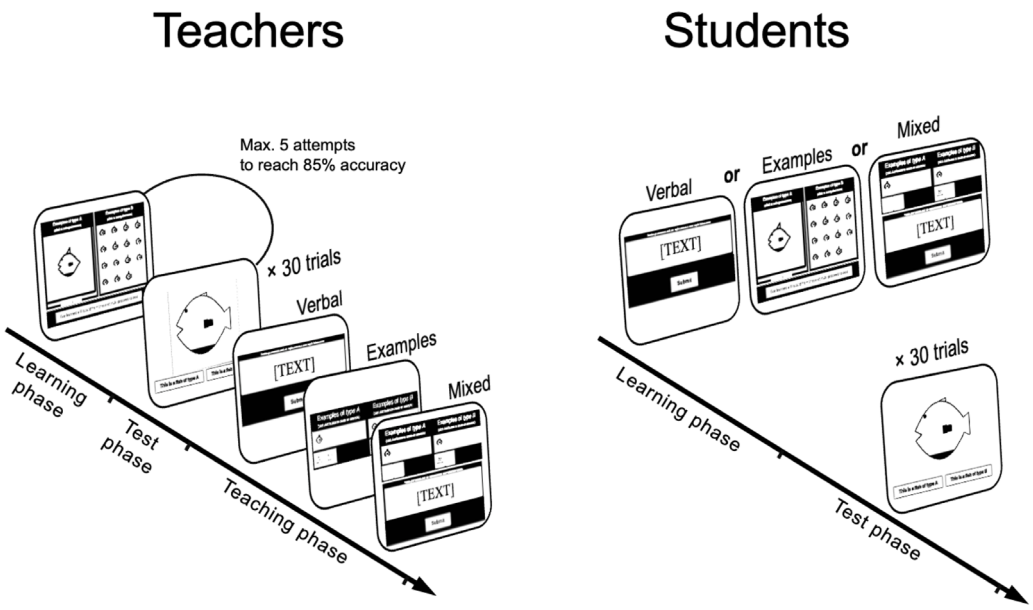

Fig. 1. Experiment 1 procedure illustration. Note that every teacher generates three types of teaching materials (for different students), but each student only receives one type.

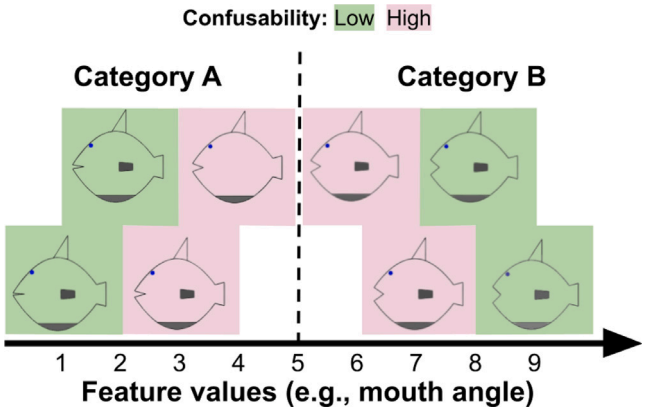

Fig. 2. Perceptual confusability illustration. The key feature in this case is how open the mouth is. In the case of high confusability, the widest open mouth in category A is close to the most narrowly open mouth in category B. In the case of low confusability, there is a larger gap.

Communication volume metrics. Looking only at student accuracy is limiting because two conditions may result in equal student performance while requiring different amounts of communication to reach that performance. Such a result would still be important, hence we also looked at communication volume in different conditions. To quantify communication volume, we stick to the most natural approach, namely using the number of words and the number of examples for verbal and exemplar-based channels respectively.

To create a single measure of communication volume for the mixed condition, however, one would need to determine a conversion procedure between examples and words. Choosing the most principled way to do that opens a methodological can of worms, hence we decided to avoid it. Instead, we focused on studying communication volume within isolated channels; for the mixed channel, we only looked at the typical proportion of communication volume of each modality (words and examples) compared to isolated channels.

2.1.4. Materials

The stimuli were schematic images of fish with up to four independently varying visual features (see a detailed description in Rosedahl and Ashby (2018)): mouth angle, dorsal fin height, tail height, and belly color. There were nine possible values within each of the dimensions.

We randomized over physical instantiations of stimulus dimensions to control for potential effects of feature salience. For example, if the task involves one relevant dimension (d) and two irrelevant (n_1, n_2), for one participant, these dimensions may be "d – tail fin, n_1 – belly color, n_2 – mouth angle", while for another, they may be "d – dorsal fin, n_1 – tail fin, n_2 – belly color". These random assignments were kept fixed between any given teacher and their students.

2.1.5. Participants

We recruited 123 teachers and 345 students via Amazon Mechanical Turk. Teachers were compensated at a base rate of $1, with an additional bonus of $1 if they reached and completed the teaching stage. Students were compensated $0.25 with a bonus of $0.25 if they reached an accuracy of 0.75. We excluded ten teachers: three of them did not reach the accuracy criterion, seven more failed to provide adequate verbal instructions. Importantly, here we do not refer to poorly phrased or low-quality instructions, but rather (1) nonsensical instructions (e.g."I think is good achieve goal"), (2) instructions that demonstrate fundamental misunderstanding of the task (e.g. when a teacher clearly assumes that a student in the verbal condition would also see examples along with their explanation: "Study the examples and guess the rule"), and (3) instructions that could not conceivably teach how to perform the task (e.g. "guess correctly the answers").

We also excluded nine students that received materials from previously excluded teachers and 24 more students who had accuracy below two standard deviations from the mean (lower than 32%). Such an accuracy is substantially below chance, meaning that these students had most likely misunderstood their teacher, learning a rule opposite to the actual one. Six students who indicated that they have poor knowledge of English or did not respond to the question were excluded as well. The final sample consisted of 113 teachers and 316 students.

2.2. Results

Our main hypotheses are concerned with how well category communication would work under different channels, and how this communication would be affected by confusability and dimensionality. These main questions are considered in 2.2.1 and . Section 2.2.2 provides supporting analysis, looking at the volume of teacher's messages, in order to better understand the mechanisms behind the main results. This

additional analysis was not, however, exploratory, since our hypotheses for these additional sections were directly informed by the results of our pilot study (see Appendix B), and in Experiment 1, we aimed to replicate these results, rather than find anything new.

2.2.1. Student performance

Student performance is, ultimately, the most important measure of whether category communication was successful. Students' accuracy was relatively high across all experimental conditions, (Mdn = .833, IQR[.567, .975][1]), showing that participants were generally able to communicate category knowledge. Mixed communication showed best results (Mdn = .90, IQR[.683, 1.0]), outperforming both example-based (Mdn = .833, IQR[.567, .967]) and verbal (Mdn = .783, IQR[.533, .967]) channels, suggesting that combining different modes of communication gives an advantage in category communication. Additionally, students in the low confusability condition (Mdn = .967, IQR[.567, 1.0]) outperformed those in high confusability (Mdn = .80, IQR[.533, .900]).

Student accuracy, however, may reflect differences in teacher's category mastery rather than communication quality. To visualize communication success across all conditions while accounting for this fact, we used a simple "accuracy loss" metric, equal to the student's accuracy subtracted from her corresponding teacher's accuracy. Under perfect communication, accuracy loss should be close to zero, while large numbers would indicate failure to communicate knowledge. On Fig. 3, we can clearly see that mixed condition results in lower accuracy loss. Additionally, we can see that higher confusability was associated with higher accuracy loss for all communication channels.

For statistical analysis, we regressed student accuracy onto learning format (verbal, examples, mixed), confusability (low or high), and stimulus dimensionality (two, three, or four), using a binomial regression model (that is, a GLM with a logistic link function and a binomial random component), and doing statistical inference using robust variance estimation. To control for teacher performance, we added a logit of teacher accuracy as a predictor. Unless otherwise specified, communication format was dummy-coded, using mixed communication as the base level. The number of dimensions was coded as a linear predictor. Lastly, confusability was dummy-coded with "high" as the base level. See Appendix C for additional information on our analysis approach.

We observed that low confusability led to better performance ($\beta = 0.46, p = .007$), and that the mixed channel outperformed both the exemplar ($\beta = -0.48, p = .006$) and the verbal ($\beta = -0.55, p = .002$). Notably, if one of the isolated channels was re-coded as the base level, no significant difference was obtained between verbal and exemplar channels ($\beta = .07, p = .248$). Other predictors were not significant. After significant main effects were established, we used the effect-coded version of the model above to test the interaction between confusability and channel. None of the interaction coefficients were, however, significant. For complete coefficient information, see Appendix G.

Bayesian analysis. Statistical analysis approach used above is standard for when accuracy is the target variable, but, as we can see on Fig. 3, performance distribution in every condition is noticeably bimodal, which makes this approach suboptimal. Since we were using robust variance estimation, this bimodality does not automatically render our inferences invalid (the mean model may still be correct, despite the misspecified variance), but the model certainly takes an oversimplified view of the data, potentially being unable to capture its key properties (one can find vastly different two-peaked distributions with the same overall mean; for the binomial GLM, these configurations are indistinguishable).

A likely explanation for why we see such a distribution is that a student either succeeds in understanding the gist of the communicated message and gets into the high-performing group (accuracy loss near zero), or fails to understand anything and performs at chance (accuracy loss near 0.5). A Bayesian mixture model is a natural choice for statistical analysis of such data.

We modeled student performance in each condition as a mixture of two distributions: the high-performing subgroup and the communication failure subgroup (performing at chance). Thus, every condition had two variables associated with it: (1) Probability of successful communication. (2) Accuracy in the successful subgroup, i.e. the probability of giving a correct answer in the case of successful communication. We then estimated the effect of each experimental variable on these probabilities, separately for verbal and exemplar-based channels (see further detail on the model in Appendix D).

Results are presented in Table 1. In accord with what can be visually seen on Fig. 3, successful student subgroups in all conditions were negatively affected by high confusability. That is, even when communication was generally successful, high confusability made it hard for students to reach mastery. At the same time, in the verbal channel higher confusability also led to an increased risk of a complete communication failure. This result agrees with our initial hypothesis about verbal communication being less suitable for situations requiring nuanced perceptual distinctions.

Student performance summary. Overall, this section presented key results of the first experiment, highlighting the differences between the verbal, exemplar, and mixed communication channels and how they are affected by our interventions. As we expected, mixed communication led to better performance (although at the cost of substantially higher volume, which we did not anticipate prior to the pilot experiment). Additionally, while all channels were negatively affected by perceptual confusability, this was especially true for verbal communication. At the same time, stimuli dimensionality did not significantly affect performance, contrary to our expectations. The next section presents an additional communication volume analysis that helps to interpret these results. It should be seen as secondary (it was not, however, exploratory, as all hypotheses tested there were directly informed by our pilot study, see Appendix B).

2.2.2. Communication volume

Teachers were free to choose how many materials (words or examples) they generate, we refer to this as *communication volume*. Table 2 shows the median numbers of words and exemplars teachers created when using isolated and mixed communication channels. Notably, in the mixed channel, teachers communicated 77% as many words as in the isolated verbal channel, and 100% as many examples[2]. Although adding those numbers can only be done with caution since 100% of verbal volume might, in subtle ways, not be exactly equivalent to 100% of exemplar volume, it is clear that teachers generated much more materials overall (nearly a sum of materials in isolated channels). This suggests that either teachers are naturally more motivated to produce diverse teaching materials, or that teachers believe that producing more materials within isolated channels results in diminishing returns.

To see whether "more is better" when it comes to teaching materials within isolated channels, we ran a subsample analysis, adding communication volume as a predictor, while still controlling for study conditions. Thus, for the exemplar channel subsample, we regressed student accuracy on confusability, dimensionality, teacher performance, and the number of examples; for the verbal channel subsample, we regressed student accuracy on confusability, dimensionality, teacher performance, and the number of words; mixed subsample was excluded

[1] "IQR" stands for Interquartile Range, and is reported in the format [a, b], where a is the 25th quantile, and b is the 75th quantile.

[2] Note, however, that although the quantiles for exemplar-based communication volume coincide, the average number of examples in the isolated examples channel (5.08) is higher than that in the mixed (4.2).

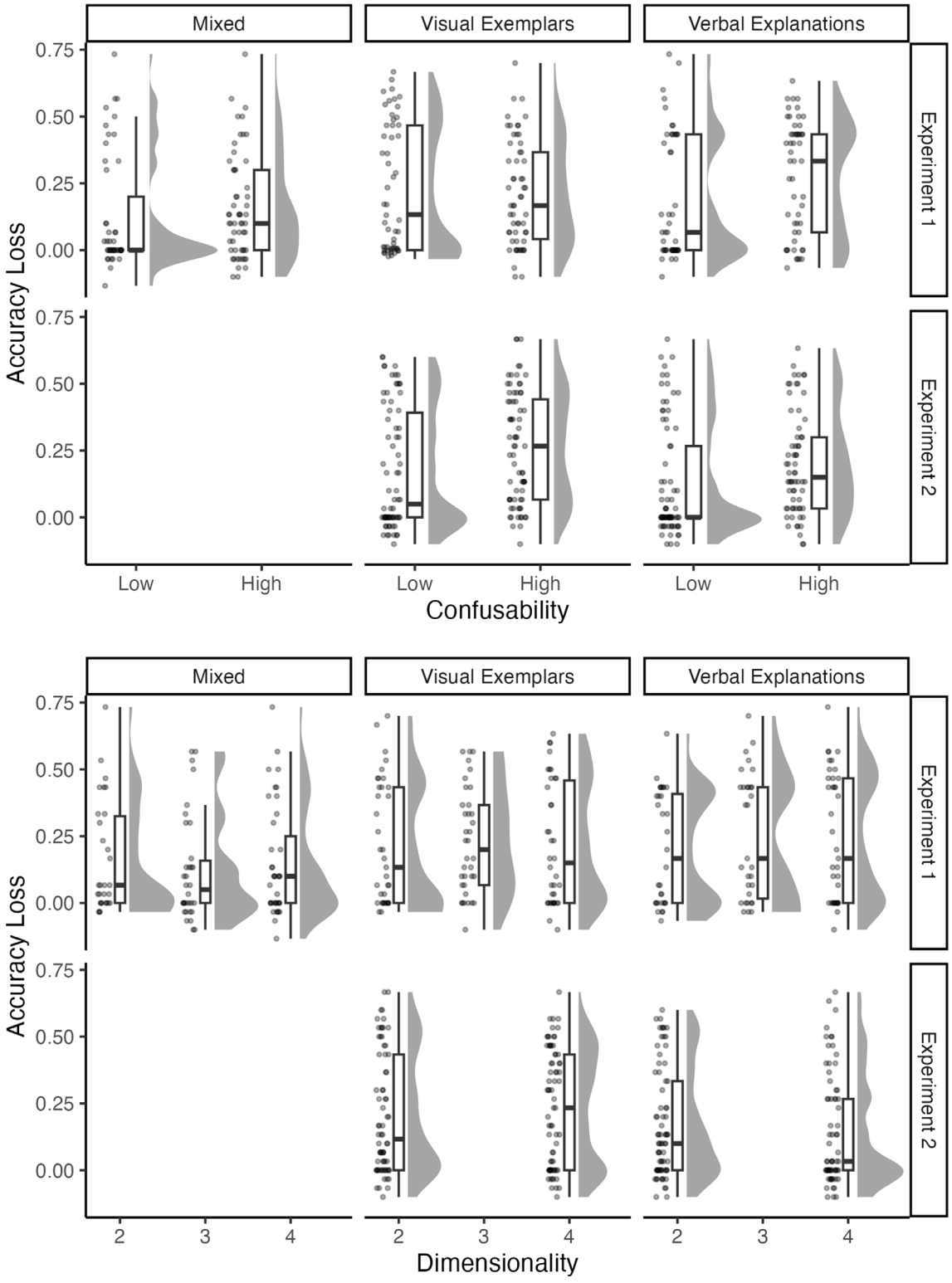

Fig. 3. Accuracy loss (difference between teacher's and student's accuracies) for all conditions in Experiments 1 and 2. Low values indicate successful knowledge transmission. High values indicate failure to communicate category knowledge. Values below zero mean that students outperform their teachers.

Table 1
Credible intervals for the impact of conditions on student accuracy and on the probability of successful communication in Experiment 1, split by communication channel.

Channel	Independent variable	Change in the probability of learning 95% c.i.	Change in accuracy 95% c.i.
Verbal	Confusability	(−0.819, −0.047)[a]	(−0.198, −0.073)[a]
	Dimensionality	(−0.396, 0.378)	(−0.092, 0.035)
Examples	Confusability	(−0.338, 0.406)	(−0.235, −0.120)[a]
	Dimensionality	(−0.386, 0.354)	(−0.064, 0.051)
Mixed	Confusability	(−0.574, 0.171)	(−0.243, −0.120)[a]
	Dimensionality	(−0.183, 0.563)	(−0.029, 0.093)

[a]Strong influence (two-sided 95% credible interval does not overlap with zero).

Table 2
Median (and IQR) number of words and exemplars communicated by teachers through different channels in Experiment 1.

	Communication channel		
	Verbal explanations	Visual exemplars	Mixed
Number of words	26 (17–42)	–	20 (11–32)
Number of exemplars	–	4 (2–6)	4 (2–6)

from this analysis. The number of words was not significantly associated with student accuracy ($\beta = 0.00, p = .967$), and the number of examples was negatively associated with student accuracy ($\beta = -0.04, p = .005$)[3].

Overall, simply increasing communication volume within isolated channels does not lead to improved performance. Therefore, it seems likely that mixed communication has an advantage not because of simple redundancy or participants' higher motivation to generate diverse materials, but because verbal and example-based channels communicate different aspects of category knowledge.

At the same time, as can be seen on Fig. 4, teachers generally produced more materials (words/examples) in more difficult conditions (high confusability, high dimensionality). To statistically test whether dimensionality and confusability affected communication volume, we excluded the mixed condition (since it involves both verbal and exemplar communication which severely complicates total production volume analysis) and separately analyzed the "verbal" and "exemplar" conditions using simple rank methods. In the exemplar channel, both confusability (Wilcoxon $W = 2017.5, p = .012$) and dimensionality (Kendall $\tau = 0.2, p = .033$) were significantly associated with a higher number of generated visual examples. In the verbal channel, higher confusability was only marginally significantly associated with a higher number of words ($W = 1894, p = .084$), while the number of dimensions had no effect ($\tau = -0.05, p = .633$).

Overall, our subsample analyses suggest that teachers adjusted communication volume depending on condition difficulty, but that increasing communication volume beyond some adequate level for a given condition does not result in improved performance.

2.3. Experiment 1 summary

The primary goal of this experiment was to establish whether the choice of the communication channel (verbal, exemplar, or mixed), and category structure (confusability and dimensionality) affected category communication. We were especially interested in whether communication channels may be differentially affected by certain aspects of category structure. In particular, we expected that verbal communication would be more severely affected by high confusability, while exemplar channel would be more affected by changes in dimensionality. Additionally, we expected that the mixed channel will be more efficient (per communication unit).

First, the data strongly suggest that the mixed channel is more effective (but not necessarily more efficient). Both in this experiment and in our pilot study, mixed communication showed better results (higher student performance) than isolated channels. At the same time, teachers generated substantially more materials overall when using the mixed channel. Therefore, we cannot say that the mixed channel is more efficient per communication unit.[4] That being said, communication volume analysis 2.2.2 suggests that the increase in effectiveness of the mixed channel is not due to sheer redundancy, but rather that verbal and exemplar channels are focused on different aspects of category knowledge.

Second, we found that category structure indeed affected category communication. In particular, higher confusability made it harder to communicate knowledge, and, in the case of verbal communication, made complete communication failure (i.e. situations when students learn nothing from their teachers and perform at chance) more likely. This is in accord with our hypothesis that verbal communication will be more dramatically affected by perceptual confusability. At the same time, we found no effect of dimensionality.

3. Experiment 2

In our second experiment, we introduced strict limits on communication volume. By doing so, we hoped to highlight the differences between communication channels. The rationale behind this hope was twofold.

First, in our pilot experiment (Appendix B) and in Experiment 1, teachers adjusted the volume of their messages to counteract the study interventions, generating more materials in difficult conditions. Thus, even in conditions where communication was difficult (as indicated by higher communication volume), it was still effective (students were able to achieve relatively high accuracy). With communication volume restricted, teachers would not have the option to adaptively change it and thus potentially counteract the effects of condition difficulty. Second, in Experiment 1 and in the pilot, students often showed near-ceiling performance, which might have masked some of the effects. Limiting communication volume makes it more challenging and thus

[3] This effect becomes non-significant if unusually high numbers of examples (more than 10) are clipped ($\beta = -0.04, p = .268$) or excluded from the analysis ($\beta = 0.00, p = .960$). Importantly none of the strategies shows evidence for the "more is better" effect.

[4] It is important to note that using other metrics of efficiency can give different results. Specifically, in our experiment, students were noticeably faster in the verbal condition. See Appendix F for details.

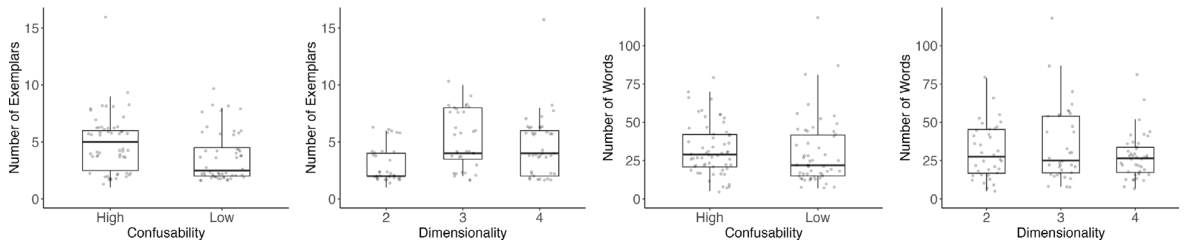

Fig. 4. Communication volume (i.e., the number of visual exemplars and words produced by teachers) in Experiment 1 (for exemplar and verbal channels). Three cases with exceptionally high numbers of exemplars (20 or more) were excluded to enhance plot readability.

can help avoid ceiling effects, increase meaningful variation in student performance, and highlight the effects of our study interventions.

The procedure followed Experiment 1, except for the exclusion of the mixed condition. Enforcing a limit on the total communication volume in the mixed condition would have involved explaining "example-to-word" conversions to participants, severely complicating the procedure. As in Experiment 1, we manipulated perceptual confusability and stimulus dimensionality, but this time we only included extreme values of the stimuli dimensions variable (two and four dimensions).

3.1. Method

3.1.1. Conditions

Teachers had two independent variables: confusability (high vs low) and stimuli dimensionality (2 or 4). Students had one additional independent variable (mode of learning): verbal or exemplar-based.

Communication volume was restricted to 2 examples and 10 words in the exemplar-based and verbal conditions respectively. These numbers were chosen based on previous experiments. Specifically, we looked at the easiest condition (low confusability, low dimensionality), and picked the *minimal* number of words and examples that resulted in successful communication as our communication volume limits. These numbers were 2 examples and 10 words respectively.

It is worth mentioning that this scheme was slightly more restrictive towards verbal communication. For verbal communication, 10 words was an unusually small number (there was only one successful teacher with such a short message). The 25th percentile for verbal message length was at 15.75 words, while the median number was 21.5. At the same time, for exemplar-based communication, using only 2 examples was typical and coincided with the median number of examples.

3.1.2. Procedure

The procedure mirrored that in Experiment 1, with the only difference that during the teaching stage, there was a limit on the number of words and examples that teachers were allowed to generate, and there was no mixed condition.

3.1.3. Participants

We recruited 108 teachers and 311 students via Amazon Mechanical Turk. Teachers were compensated at a base rate of $1, with an additional bonus of $1 if they reached and completed the teaching stage. Students were compensated $0.25 with a bonus of $0.25 if they reached an accuracy of 0.75. Pre-defined accuracy criterion of 85% was reached by 85 teachers, the rest were excluded from further analysis. Fourteen teachers failed to provide adequate verbal instructions and thus were excluded as well (using the same procedure as described in 2.1.5). We also excluded 20 students: 18 of them had categorization accuracy below two standard deviations from the mean (below 20%) and two students indicated poor knowledge of English. The final sample of teachers consisted of 71 teachers and 291 students.

3.2. Results

3.2.1. Student performance

Despite the volume restrictions, category communication was, overall, successful: median student accuracy was .87 (IQR[.567, 1.0])[5].

We, again, observed strong bimodality, and hence opted for a Bayesian mixture model as our main analysis instrument. When we apply this model, first, we see in Table 3 that confusability negatively affected accuracy in successful subgroups (both in exemplar-based and in verbal communication). Second, the probability of communication (getting into the successful subgroup) in the exemplar-based condition was negatively affected by both confusability and dimensionality. Both effects are borderline on 0.95 two-sided level, but significant if a one-sided interval is used.[6] At the same time, the probability of successful verbal communication is not significantly affected neither by confusability nor by dimensionality, although in the case of confusability, there seems to be a trend suggesting that a weaker effect is potentially present, which is especially likely, considering the results of the first experiment.

Overall, the second experiment shows that verbal communication is more resilient in situations of restricted communication volume, especially against high stimuli dimensionality. This resilience against dimensionality is in line with our original hypothesis about the different specializations of verbal and exemplar-based communication channels, although initially, we expected it to be true universally, and not only in restricted communication volume conditions.

One seemingly counterintuitive result warrants a separate mention: increasing dimensionality positively affects accuracy within the successful subgroup in the case of exemplar-based communication. A likely explanation is that higher dimensionality makes it harder to communicate the concept, but does not severely affect concept application if the communication is successful. Indeed, when one learns which features to look for, other features can be easily ignored, but it might be difficult to identify/communicate relevant vs irrelevant dimensions initially. Thus, if the communication is successful, dimensionality does not dramatically affect performance. Naturally, in the high dimensionality condition, only the more motivated or talented teacher–student pairs make it to the successful subgroup. They show better results than a successful subgroup in a low dimensionality condition, which, due to the ease of communication in that condition, includes a mixture of students of different levels of motivation and ability. In short, dimensionality seems to affect learning the concept, not its application, hence the "communication success" group under high dimensionality condition is formed by more talented/motivated participants, who perform slightly better.

The effects of confusability, in contrast, do not exhibit such a pattern. A likely reason is that confusability not only affects the difficulty

[5] Note that, surprisingly, the median value is slightly higher than in Experiment 1 (.833); we will return to this fact in general discussion.

[6] If suspiciously fast students are removed from the dataset (see Appendix E), the effect of dimensionality strengthens and becomes significant, while the effect of confusability weakens, losing even marginal significance.

Table 3
Credible intervals for the impact of conditions on student accuracy and on the probability of successful communication in Experiment 2, split by communication channel.

Channel	Independent variable	Change in the probability of learning 95% c.i.	Change in accuracy 95% c.i.
Verbal	Confusability	(−0.498, 0.110)	(−0.211, −0.111)[b]
	Dimensionality	(−0.397, 0.213)	(−0.036, 0.063)
Examples	Confusability	(−0.584, 0.041)[a]	(−0.191, −0.096)[b]
	Dimensionality	(−0.611, 0.013)[a]	(0.003, 0.097)[b]

[a]Moderate influence (two-sided 95% credible interval overlaps with zero, but a one-sided does not).
[b]Strong influence (two-sided 95% credible interval does not overlap with zero).

of concept communication but also the difficulty in applying the concept, even after it was successfully communicated. Hence the successful subgroup, although consisting of slightly more motivated individuals, still experiences a drop in performance in the high confusability condition.

3.2.2. Comparing performance in experiments 1 and 2

To see how volume restriction affected performance, we ran additional analysis by combining data from Experiments 1 and 2, filtering out the mixed condition, and adding the experiment indicator variable (with Experiment 1 as the base level) to our main GLM model (accuracy as target, and confusability, communication channel, dimensionality, and logit of teacher accuracy as predictors). The overall effect of experiment was marginally significantly positive ($\beta = 0.189, p = 0.083$), meaning that restricted volume, paradoxically, led to better results. Further examination revealed that the effect was entirely driven by the verbal subgroup. In the example-based communication subgroup, the effect went away ($\beta = 0.01, p = 0.949$). In the verbal subgroup, in contrast, the effect was strong and significant ($\beta = 0.396, p = 0.013$). That being said, mixed communication in Experiment 1 still outperformed both isolated channels in Experiment 2 ($\beta = -0.33, p = 0.031$).

3.2.3. Communication volume

Most teachers used all or almost all available communication volume to communicate their knowledge. Exemplar-based communication channel showed no variability at all, with all teachers using 2 examples in all conditions. For the verbal channel, there was a marginal variation with the median ranging from 9 to 10 across all conditions.

3.3. Summary

When the amount of communication is restricted, we see a qualitatively different pattern in how communication channels are affected by confusability and dimensionality. Verbal communication was more robust when it comes to ensuring that at least some useful information was communicated. The most pronounced difference was the way in which communication channels reacted to changes in stimulus dimensionality: verbal communication was unaffected by this factor, while exemplar-based communication became problematic. Specifically, under high stimulus dimensionality, there was a high risk that exemplar-based communication will fail entirely.

This effect of dimensionality on communication effectiveness presents an interesting contrast with the first experiment and the pilot. Previously, stimuli dimensionality did not affect student accuracy but affected communication volume. Now, when communication volume is fixed, we see the effect on accuracy, which supports the idea that the influence of irrelevant dimensions can be compensated by increasing communication volume, at least in the context of exemplar-based communication. We also see that under a restricted volume scenario, stimuli dimensionality affects exemplar-based communication more than verbal communication.

4. Discussion

Real-world knowledge transmission is greatly aided by the use of language, but there is little research on language as a means for *category* communication. To bridge this gap, we conducted two experiments studying verbal, exemplar-based, and mixed-channel category communication. We were especially interested in the differences between these modes of communication. In this discussion, we revisit the major hypotheses of our study, discuss whether they were supported by data, and consider the implications of our results.

The first universal result that we observed was directly tied to one of the main hypotheses of our study. We expected that combining different communication channels would be the most efficient (per unit volume) way of transferring category knowledge. Indeed, both in our pilot experiment and in Experiment 1, mixed communication (when teachers were allowed to communicate both verbally and by generating exemplars) led to *superior student performance*. At the same time, mixed communication was also associated with dramatically higher communication volume. Therefore, while being the most effective, it was not the most efficient way of communication,[7] rendering the hypothesis only partially confirmed (or partially disproved, depending on one's outlook).

We see two potential reasons for the effectiveness of mixed communication. On the one hand, it is possible that not all knowledge can be reliably transferred via isolated channels, hence when communication is restricted to a single channel (verbal or exemplar-based), some information is lost. On the other hand, people generate more materials overall in the mixed condition (compared to isolated channels); therefore, it may be that teachers communicate more successfully in the mixed condition simply through redundancy. The latter seems less likely for two reasons. First, we observed no evidence that higher communication volume leads to higher student accuracy when controlling for experimental conditions. That is, more materials were not always better, and, therefore, simply providing redundant information in the same channel would not allow to catch up with the effectiveness of mixed communication. Second, communication volume in Experiment 1 was not restricted, i.e. teachers were free to generate more materials in isolated channels, but apparently did not believe that doing so would help their students. Overall, it seems more likely that mixed communication is advantageous because verbal and exemplar-based communication are tailored to different aspects of category knowledge. In the context of existing literature, previously, it had been shown that verbal descriptions can help category learning when explicitly linked to specific regions/dimensions of the stimulus (Miyatsu, Gouravajhala, Nosofsky, & McDaniel, 2019). We expand this result by showing that a mixture of verbal and exemplar-based communication generally outperforms communication via isolated channels. More broadly, this result shows that exemplar-based and verbal channels are not interchangeable, which provides indirect support for theories that postulate that

[7] It is important to note that we are specifically speaking about efficiency as student accuracy per unit volume. Focusing on efficiency per unit time may give different results. See Appendix F for further discussion.

language has a unique role in category learning (Ashby et al., 1998) and more generally (Clark, 1998; Gentner, 2016; Lupyan, 2012), as opposed to being viewed as an equivalent way to re-code perceptual experiences.

Our next major hypothesis was that language should be particularly effective when communicating categories with a high number of irrelevant dimensions because it enables focusing attention on the most informative characteristics of stimuli (Kloos & Sloutsky, 2008; Sloutsky, 2010). However, in Experiment 1, we found no differences in how verbal and exemplar-based communication performance was affected by stimuli dimensionality. We suspected that the absence of effect might be due to teachers' adjustment of communication volume according to the demands of specific conditions and due to ceiling effects in student performance. In Experiment 2, we addressed these issues by restricting the communication volume. In restricted volume conditions, higher dimensionality indeed reduced the category communication effectiveness of the nonverbal channel while having no effect on verbal communication. This result is consistent with the notion that language may play a role in dimensionality reduction when teaching categories.

The last major hypothesis was that we expected high confusability to have a stronger negative impact on verbal communication, compared to exemplar-based teaching. In Experiment 1, we saw that this was indeed the case. While all modes of communication were negatively affected by confusability, the effect was more prominent for verbal communication. Specifically, only in the verbal communication channel, higher confusability significantly increased the risk of complete communication failure. There is, however, a surprising caveat to this observation: in Experiment 2, the effect got weaker (same direction, but not significant). While it may be just random variation in our samples, certain results suggest that something more insightful might be going on. Specifically, the most surprising aspect of Experiment 2 was that student accuracy went *up*, compared to Experiment 1, defying communication volume restrictions. Closer examination (see 3.2.2) revealed that this marginally significant effect was entirely driven by the verbal communication subgroup (for which it was significant at $p = 0.013$). It is possible that we see consequences of a metacognitive failure, where teachers *think* that they help their students by providing longer verbal instructions, while in fact, it is better to be forced to provide a concise and short explanation. We believe that this unexpected result warrants further exploration in future research.

We believe that in order to provide a deeper theoretical account of the observed differences between verbal and exemplar-based communication (and, hopefully, to formally explore the exact properties of category structure they are best attuned to), it is necessary to develop a computational model of the process. In this paper, we only aimed to provide empirical support for the presence of qualitative differences between the channels and gain a high-level understanding of what these differences are. Although developing a computational model is outside of the scope of this paper, we would like to mention a few directions that could provide a starting point for such modeling. One would be to build upon a prototype that was suggested in Moskvichev et al. (2019); the authors expanded the model by Shafto et al. (2014) by adding a high-level account of verbal communication. That model aims to capture which categories, generally, are better suited for verbal communication, but does not make any predictions about specific words that participants might use. An alternative approach would be to develop a more explicit model of category learning from language that would be capable of learning categories from natural language texts. Such prospects become realistic due to the advance of neural Natural Language Processing architectures (Brown et al., 2020) that, after pre-training on a large corpus, can be fine-tuned to novel tasks with relatively small amounts of data (Malte & Ratadiya, 2019) and can be adapted to model learning after the initial training stage is over (Hutchins, Schlag, Wu, Dyer, & Neyshabur, 2022; Moskvichev & Liu, 2021).

Apart from the specific hypotheses covered above, it is important to mention one prominent result present in all experiments: the general *robustness of category communication*. Teachers were able to successfully communicate their knowledge in all conditions, even when communication volume was severely restricted. This result expands two previous lines of Cognitive Science research. On the one hand, Avrahami et al. (1997) and Shafto et al. (2014) showed the benefits of learning categories through exemplars generated in a pedagogical, rather than random fashion. At the same time, Chopra et al. (2019) showed that verbal category communication can be effective (students get accuracy close to that of their teachers), but provided no direct comparison with exemplar-based communication. Our results are the first to directly compare exemplar-based and verbal category communication and to establish that they result in similar (although not equivalent) and generally high performance across a wide range of conditions. One of the most important conclusions of our study is that verbal communication provides a viable way of category learning but with markedly different dynamics from that of learning by examples. Given how common language-based category acquisition is in practice, we hope that it becomes one of the standard ways of studying category learning in laboratory settings, continuing the general trend towards higher environmental validity of category learning studies (with realistic stimuli now used more often (Nosofsky, Sanders, Gerdom, Douglas, & McDaniel, 2017; Rosedahl & Ashby, 2018), and more attention paid towards studying category in situations where the source of information is not a neutral environment, but a knowledgeable "other" (Shafto et al., 2014)).

Our study has a number of limitations that are important to mention. The most substantial limitation is the narrow range of category structures we considered; in our experiments, we focused on a family of simple one-dimensional rules. In the pilot experiment, we also used a two-dimensional rule, but again, with a very simple structure (a conjunction of two one-dimensional rules). Such rules may be better suited for language-based communication than, for example, *information integration* category structures (Ashby et al., 1998; Ashby & Maddox, 2005; Minda & Miles, 2010; Rosedahl, Serota, & Ashby, 2021). Although such simplifications were necessary to keep the scope of the study manageable, we believe that in the future, it is important to expand the range of category structures. That will allow us to better understand the benefits and limitations of different modes of pedagogical communication in category learning.

Another limitation is that we do not collect field data on the frequency of verbal category communication in real-life situations (e.g. when a mother teaches a new concept to her child, or when a teacher presents new material). We do see that when mixed (verbal and exemplar-based) communication is allowed, teachers do use both communication channels, showing that people often choose to communicate categories verbally, at least in a laboratory setting. Nevertheless, we believe that collecting more naturalistic data on category teaching behavior would be highly beneficial and should be done in the future.

5. Conclusion

There has been a push for studying category learning in situations with more realistic and higher-dimensional stimuli, as well as in pedagogical (teacher–student) rather than neutral (environment-student) scenarios. Building upon the previous results, our study makes the next step by focusing on language-based category communication, which is common in day-to-day category acquisition but is rarely studied.

Theoretically, we establish a number of ways in which exemplar-based, verbal, and mixed communication differ from each other. Specifically, we saw that mixed communication was the most effective, and that verbal and exemplar-based communication may be tailored towards slightly different situations. Verbal communication was better suited for quick communication of the gist of the category knowledge and being resilient against changes in stimuli dimensionality,

while exemplar-based communication, in turn, was more robust in situations with no restrictions on communication volume and when high perceptual precision was required. Most importantly, these results show that language provides a viable but not equivalent alternative to exemplar-based category communication, with its own unique dynamics. It yields indirect support to theories that postulate a unique role of language in category learning. Indeed, under theories that see language as reducible to an equivalent way of receiving the same information, or as the inevitable universal final step in knowledge representation, we should expect no difference between exemplar-based and language-based category communication.

On the practical side and outside of the scope of Cognitive Science, our results provide a controlled illustration of the importance of using both verbal explanations and examples in real-life teaching situations (e.g. school or college). Additionally, our results might help in developing robotic and Human-Computer Interface systems. For example, efforts on combining verbal and nonverbal instructions for artificial intelligence systems, such as Yu and Mooney (2022) and Li et al. (2019), might benefit from a deeper understanding of the advantages and limitations of using language to transfer knowledge.

We hope that the methodology that we developed and the results that we obtained will serve as a foundation for further research on the role of language in category communication, and, more generally, in understanding how humans share knowledge via language.

CRediT authorship contribution statement

Arseny Moskvichev: Conceptualization, Methodology, Software, Formal analysis, Investigation, Data curation, Writing – original draft. **Roman Tikhonov:** Conceptualization, Methodology, Software, Formal analysis, Investigation, Data curation, Visualization, Validation, Writing – review & editing. **Mark Steyvers:** Conceptualization, Methodology, Supervision, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data and analysis for this paper are publicly available at https://osf.io/hjyu5/.

Acknowledgments

We are grateful to Alexey Kotov for useful suggestions on experimental procedure and to Marina Dubova for her feedback on the manuscript.

This research was partially supported by the John I. Yellott Scholar Award to Arseny Moskvichev (2019).

Appendix A. Study interface detail

Exemplar-based learning phase interface is illustrated on Fig. 5. The overall teaching interface is illustrated on Fig. 6, with the interactive slider-based window for providing examples illustrated separately on Fig. 7.

Appendix B. Pilot study detail

In this section we describe our pilot study. It had a relatively large sample, contributed a few important results regarding communication volume, and informed the hypotheses tested in Experiment 1 (except for the hypotheses that were pre-registered before the pilot).

The procedure was similar to Experiment 1, but with a few important differences. First, in Experiment 1, we changed the operationalization of perceptual confusability, making between-category distance lower than within-category variability, in the high-confusability scenario, which was done to strengthen the intervention effect. In the operationalization we used in the pilot study, high confusability only affected stimuli which were close to the category boundary, while a substantial number of exemplars were still easily classifiable. Second, in Experiment 1, we only used one-dimensional rules, while in the pilot, we also had a two-dimensional rule condition. The two-dimensional rules were simple conjunctions of one-dimensional rules. We removed this condition in Experiment 1, as it contributed to a disproportionally high dropout, complicating the analysis. When it comes to volume analysis, in the pilot experiment, we used a simple conversion procedure between examples and words to calculate "total communication volume" in the mixed condition, while in the first experiment, we avoided this conversion, focusing on simpler channel-specific analyses instead. The last difference was that, in contrast with the first experiment, teacher bonus compensation was bound to their student performance. We removed this dependency in Experiment 1 primarily due to technical and ethical reasons as sometimes students might perform poorly even if the teacher did their best to ensure reasonable performance; in the pilot experiment, we had to resolve a large number of bonus assignment cases manually.

B.1. Design: independent variables

We used a three factor between-subject design. *Teachers* were assigned into one of twelve groups defined by the following category characteristics: rule dimensionality (one- or two-dimensional rules), stimulus dimensionality (two, three, or four varying dimensions), and perceptual confusability (low or high).

Rule dimensionality. We had two levels of the rule dimensionality variable. In the one-dimensional rule condition, we used rules in the form "if $x > c$ then category A else category B", where x is the numerical value along a pre-specified stimulus dimension and c is a threshold constant. In the two-dimensional rule condition, we used a conjunction of two one-dimensional rules, i.e. "if $x > c_1$ and $y > c_2$ then category A else category B".

Note that we only used "rule-based" or "verbalizable" category types in our experiments (according to the classification by Ashby et al. (1998)). Due to the nature of information-integration ("nonverbalizable") rules, verbal communication of such rules is likely to fail entirely. We, therefore, restricted ourselves to rule-based categories. It is important to clarify that the name "verbalizable" only means that such rules can conceivably be formulated verbally (Ashby et al., 1998), and does not imply that all rules of this type are equally easy to formulate or communicate verbally. As we will see, even simple verbalizable rules provide a number of challenges and insights.

Other variables. All other variables (communication channel, perceptual confusability, stimulus dimensionality) coincided with those in Experiment 1.

B.2. Participants

All participants were English speakers from the US, recruited through Amazon Mechanical Turk. Teachers were compensated at a base rate of $1, with an additional bonus of $1 if their students reached an accuracy of 0.75. Students were compensated $0.25 with a bonus of $0.25 if they

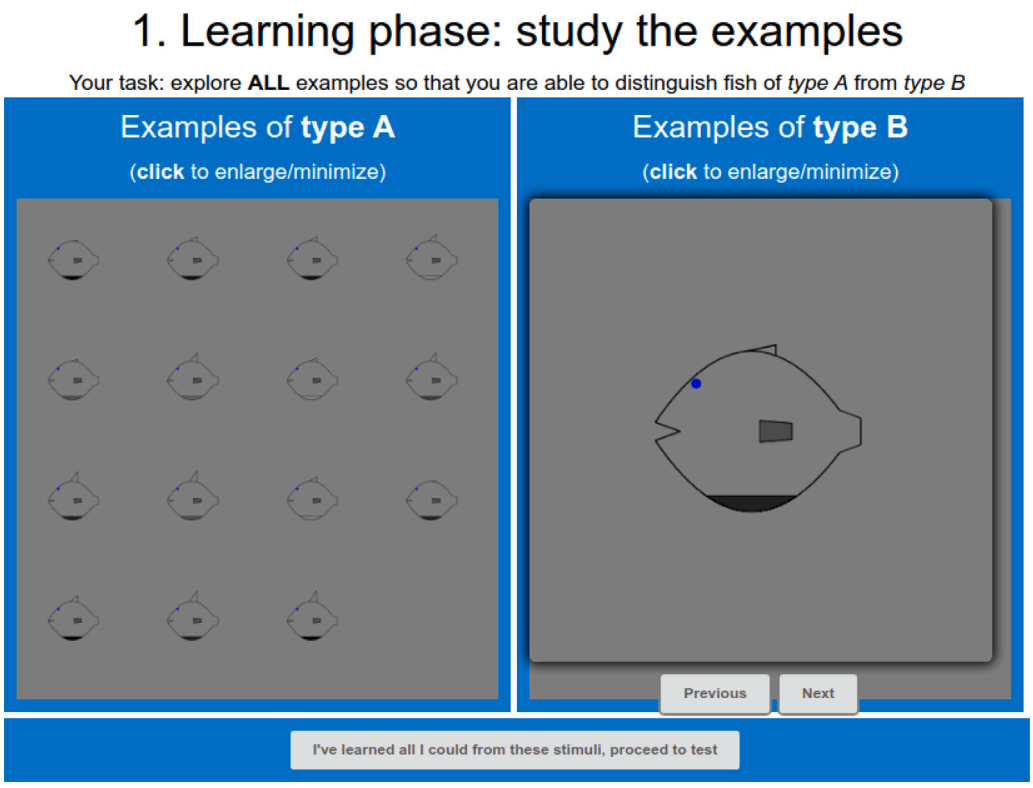

Fig. 5. Learning interface (study phase). All examples are presented at once (as seen for category A on the left), and participants are free to zoom into any given example to study it in detail (right). For teachers, there are always 15 examples randomly sampled for each category, for students, the number of displayed examples and examples themselves are generated by their respective teachers.

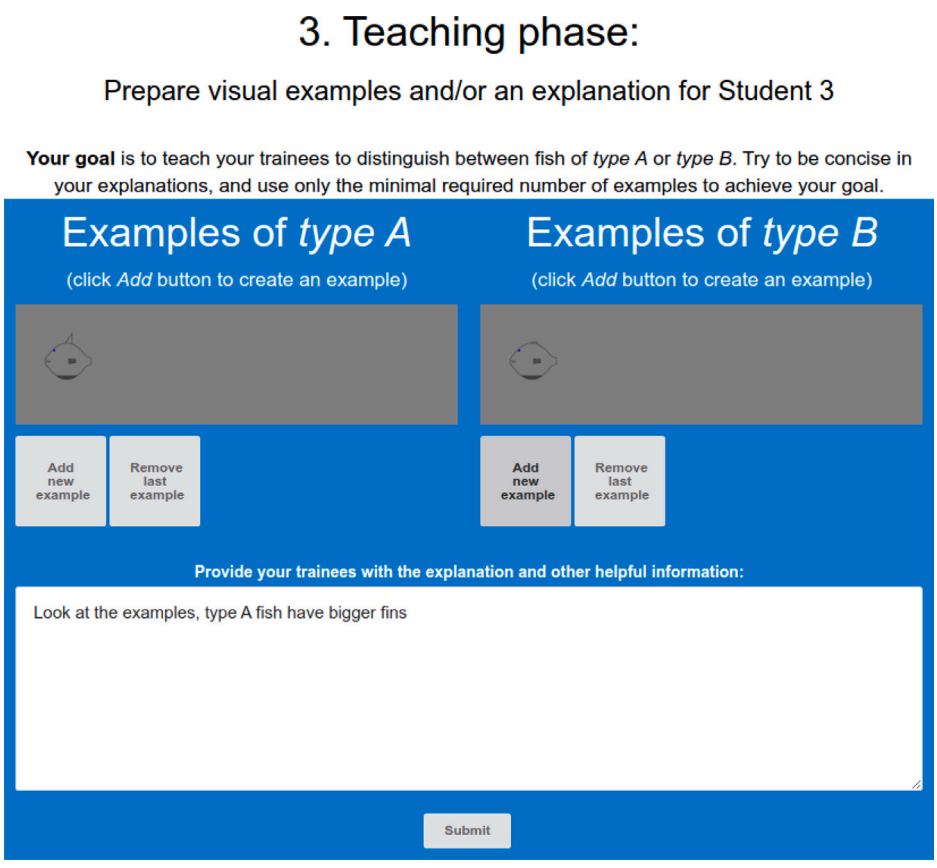

Fig. 6. Teaching phase interface for the case of mixed communication. In verbal and exemplar-based communication conditions, the interface was analogous, but with the exemplar-based and verbal textbox removed, respectively. Text input is done via keyboard, while examples are added via an interactive interface, illustrated on Fig. 7.

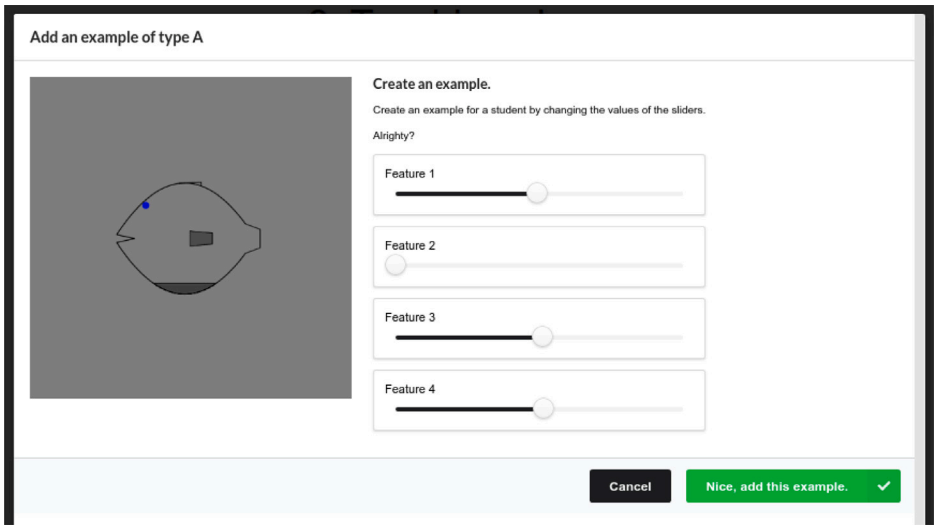

Fig. 7. Slider interface that teachers used to create examples for their students. Each slider controls one of the stimuli features (mouth size, dorsal fin size, tail fin size, belly color). In the picture, feature 2 (dorsal fin) is set to its minimal value, hence the fin on the fish's back is almost entirely gone.

reached an accuracy of 0.75. The initial sample consisted of 169 teachers and 188 students. However, we excluded 40 teachers who did not reach the predefined 85% accuracy threshold in five attempts to learn the rule. Four more teachers did not finish the experiment and were also excluded. Twenty-six teachers failed to provide adequate teaching materials (13 of them created no examples or verbal instructions and 13 provided meaningless instructions). Importantly, here we do not refer to poorly phrased or low-quality instructions, but rather (1) nonsensical instructions (e.g."I think is good achieve goal"), (2) instructions that demonstrate fundamental misunderstanding of the task (e.g. when a teacher clearly assumes that a student in verbal condition would also see examples along with their explanation, e.g. "Study the examples and guess the rule") and (3) could not conceivably teach how to perform the task (e.g. "guess correctly the answers").

Most of the excluded teachers ($n = 50$) were from the two-dimensional rule condition. Seven students with an accuracy below 2 standard deviations (37%) and nine students who received materials from previously excluded teachers were excluded as well. One student who indicated poor knowledge of English was excluded. Thus, the final analysis included 99 teachers and 171 students. The majority of teachers who were not included in the final analysis were excluded on the basis of the predefined 85% accuracy criterion (40 out of 70 excluded teachers). Students received teaching materials from a subsample of 60 teachers. Most of the students were in one-dimensional ($n = 115$) condition. The number of people in low-confusability condition was higher ($n = 107$), than in high-confusability condition ($n = 65$).

B.3. Results

B.3.1. Teacher performance

Although teacher performance is not the main focus of our hypotheses, it was still important for us to see whether our conditions affected teacher performance. First, it allowed us to test the intervention quality. That is, if teachers were to perform exactly equally across the board, it would have suggested that the category structures we consider are not sufficiently different. On the other hand, if any differences are discovered, we must account for them when analyzing student data. Otherwise, a difference in student performance between two conditions may simply be "inherited" from an analogous difference in teachers' performance, as opposed to reflecting differences in communication effectiveness.

As expected given our 85% accuracy threshold, median categorization accuracy among teachers was high, 0.97 (IQR[0.93, 1][8]). Nevertheless, some discrepancies remained: we observed slightly lower values in high confusability condition (Mdn = 0.97; IQR[0.9, 1]) and higher values in low confusability condition (Mdn = 1; IQR[0.97, 1]). Similarly, accuracy was slightly lower in the two-dimensional rule condition (Mdn = 0.97; IQR[0.90, 1]) than in the one-dimensional (Mdn = 0.97; IQR[0.97, Q3 = 1]).

Statistical analysis (binomial regression model with robust variance estimation) showed that these discrepancies were indeed significant (deviance $= 24.949, df = 3, p_{\chi^2} < 0.001$). Among individual coefficients, we observed a significant effect of one-dimensional rule type ($\beta = 0.61, p = .008$) and low confusability ($\beta = 0.87, p < .001$) on teachers' categorization accuracy. The effect of stimulus dimensionality was not statistically significant ($\beta = -0.09, p = .54$). On the other hand, we observed uneven teacher dropout (failure to reach the 85% accuracy threshold) across different stimulus dimensionality values (18.75%, 20.37%, and 29.85% for 2, 3, and 4-dimensional stimuli respectively), suggesting that stimulus dimensionality largely determined whether a categorization rule would be learned at all, but did not significantly affect the performance when the rule was successfully learned. Overall, we see that despite the high 85% accuracy threshold, the variables of interest still had an effect on teacher accuracy. On the one hand, it confirmed that the category structures in the conditions we chose were substantially different in the context of learning these categories. On the other hand, we must account for differences in teacher performance when analyzing and interpreting differences in student performance.

B.3.2. Teaching materials content analysis

To see whether teachers adjusted the content of their messages in systematic ways depending on the condition they were put in, we performed a manual content analysis. We identified seven common message types present in teacher's messages and then tagged all texts according to these types (each message may belong to more than one message type). Most messages (74%) included descriptions of typical members of each category ("Exemplars" message type). Dimensionality Reduction (explicitly stating that certain stimuli dimensions are not informative) was the second most common type at 39%. Other message types were found in less than 20% of cases each (see Table 10).

[8] "IQR" stands for Interquartile Range, and is reported in the format [a, b], where a is the 25th quantile, and b is the 75th quantile.

Table 4
Median (and interquartile range) number of words (converted to examples) and exemplars communicated by teachers through different channels. The conversion rate was calculated as the median number of examples across all teachers divided by the median number of words.

	Communication volume		
	Words (converted to examples)	Exemplars	Total
Isolated channels	4.00 (2.50–5.56)	4.00 (4.00–6.00)	8.88 (6.44-11.94)
Mixed channel	3.12 (1.69–5.00)	4.00 (2.00–6.00)	8.00 (4.38-10.34)

Table 5
Median (and interquartile range) number of words and exemplars communicated by teachers separated by communication format and stimulus type in the pilot experiment.

Stimulus type	Communication format			
			Mixed	
	Verbal	Exemplar-based	Verbal	Exemplar-based
Rule dimensionality				
One-dimensional	23.0 (18.00–34.00)	4.0 (2.00–6.00)	17.0 (10.00–32.00)	4.0 (2.00–6.00)
Two-dimensional	42.5 (30.25–53.50)	4.0 (4.00–6.00)	34.5 (20.50–47.75)	4.0 (4.00–5.75)
Perceptual Confusability				
Low	26.0 (18.50–44.00)	4.0 (3.00–6.00)	24.0 (13.00–36.00)	4.0 (2.00–5.00)
High	33.5 (23.00–45.00)	4.5 (4.00–6.50)	29.5 (14.75–42.50)	4.0 (2.00–6.00)
Stimulus Dimensionality				
Two	26.0 (18.00–45.00)	4.0 (4.00–6.00)	18.0 (13.00–36.00)	4.0 (2.00–5.00)
Three	27.0 (19.00–45.00)	4.0 (4.00–6.00)	20.0 (8.00–38.00)	4.0 (2.00–6.00)
Four	34.0 (24.00–44.00)	4.0 (4.00–6.00)	32.0 (18.00–44.00)	4.0 (4.00–6.00)

Among the seven message types, we identified two that, we expected, will be used more or less depending on our interventions. Specifically, we expected that higher stimulus dimensionality would result in a greater proportion of Dimensionality Reduction messages, while higher confusability would be associated with increased usage of Boundaries and Thresholds (to assist in distinguishing between borderline exemplars). Teachers were indeed less likely to use Boundaries and Threshold messages in the low confusability condition (5% of cases) than in the high confusability one (36%), $\chi^2(1, N = 192) = 27.76, p < .001$. However, the difference in Dimensionality Reduction between two- (34%), three- (33%), and four-dimensional (46%) conditions was not statistically significant, $\chi^2(2, N = 192) = 2.8, p = .246$. The key takeaway is that verbal messages that teachers generate differ in systematic ways, depending on condition. In the context of the pilot study, this conclusion should be treated as tentative, since we identified the categories and performed statistical analysis on the same data.

B.3.3. Communication volume analysis

Generally (as seen in Table 4), teachers produced noticeably more teaching materials in the mixed communication channel.

We also hypothesized that teachers changed communication volume based on conditions (confusability, stimuli dimensionality, rule dimensionality). More specifically, that teachers might be counteracting difficulties in communication in specific conditions by creating more materials. Descriptive statistics corroborate this idea (see Table 5). Specifically, we see that the volume of verbal communication responds to all relevant variables, increasing as the difficulty of the condition increases (i.e. when we go from one-dimensional to a two-dimensional rule, from low to high confusability, or as we increase stimulus dimensionality from two to three to four). This effect is present both in the isolated verbal channel and in the verbal component of the mixed channel. At the same time, the number of generated examples is more consistent, as the median remains exactly 4.0 in almost all conditions. Nevertheless, in the overwhelming majority of cases, the quantiles either stay the same or go up as condition difficulty increases, so it seems that the effect remains, although, potentially, less prominent.

To test this hypothesis statistically, we created a "total volume" variable. To calculate the total volume, we used a simple procedure for converting the number of communication units in one channel into another (i.e. how many examples, on average, correspond to one word). Specifically, we used a median (across all participants) number of examples in exemplar-based teaching materials and divided it by the median number of words in verbal teaching materials, obtaining a conversion constant $c_{\text{ex\_per\_word}}$. This allowed us to calculate **total information**: the amount of teaching material expressed in examples (communication units). Thus, in the exemplar-based channel, the total volume is simply equal to the number of examples. In the verbal channel, the total volume is equal to $n_{\text{words}} \cdot c_{\text{ex\_per\_word}}$. In the mixed channel, total volume is equal to $n_{\text{words}} \cdot c_{\text{ex\_per\_word}} + n_{\text{examples}}$.

After the total information was calculated, we used a gaussian glm with a log link function and robust variance estimation to evaluate the effect of stimuli characteristics (rule type, confusability, stimulus dimensionality) and teaching format (verbal, examples, mixed) on the total amount of communicated information. The log link function was chosen since the raw total information variable has a strong right skew, while its log-transformed version is reasonably close to a normal distribution. The overall model was significant ($F(5, 296) = 19.03, p < .001$). Stimulus dimensionality ($\beta = 0.09, z = 2.03, p = .043$), low confusability ($\beta = -0.2, z = -2.36, p = .018$), and two-dimensional rule ($\beta = 0.39, z = 5.17, p < .001$) were all significant predictors of total information. Exemplar-based ($\beta = -0.44, z = -3.93, p < .001$) and verbal ($\beta = -0.58, z = -8.28, p < .001$) communication conditions were also statistically significant, meaning that total communication volume was higher in the mixed condition, compared to isolated exemplar-based and verbal channels.

B.3.4. Student performance

The results in previous sections provide a general picture of teachers' communication strategies and their adaptations to different conditions. However, we need to look at student performance to see whether communication was successful.

Generally, students managed to learn categorization rules relatively well, although usually not reaching their teacher's performance. Median categorization accuracy was 93% (IQR[0.73, 1.00]) with highest value in the mixed condition (Mdn = 0.97, IQR[0.80, 1.00]) and exemplar-based condition (Mdn = 0.97, IQR[0.68, 1.00]) compared to the verbal condition (Mdn = 0.90, IQR[0.68, Q3 = 1.00]). Accuracy was noticeably lower in the two-dimensional rule condition (Mdn = 0.77, IQR[0.57, 0.93]) compared to the one-dimensional rule condition (Mdn = 0.97, [0.87, 1.00]).

For statistical analysis, we regressed student accuracy onto learning format (verbal, examples, mixed, dummy-coded with mixed as base),

rule type (one- or two-dimensional, with one-dimensional as base), confusability (low or high, with high as base), and stimulus dimensionality (two, three, or four, coded as a linear predictor), using a binomial regression model with robust variance estimation. The overall model was statistically significant (deviance $= 231.61, df = 5, p_{\chi^2} < .001$). We first identified the significant main effects: low confusability ($\beta = 0.43, p = .05$) and one-dimensional rule ($\beta = 1.14, p < .001$) both led to improved student performance. The effect of the number of irrelevant dimensions was not, however significant ($\beta = 0.08, p = .51$). Verbal communication was significantly worse than mixed ($\beta = -0.5, p = .04$), although exemplar-based communication ($\beta = -0.46, p = .09$) was only marginally worse than mixed communication (using a two-sided interval).

The results above, however, do not allow to conclude that knowledge communication is affected by intervention variables (rule-type, confusability, dimensionality). Instead, it may be that the differences in student performance simply reflect analogous differences in teacher performance. To account for that, we also fit a regression controlling for the effect of teacher performance, including a logit of the student's teacher accuracy as a predictor. The new model fit the data significantly better than the previous (deviance $= 9.276, df = 1, p_{\chi^2} = 0.002$). Under this new model, however, the weakly significant coefficients got "explained away" by teacher accuracy. Thus, only the effect of rule type remained significant ($\beta = -1.082, p < .001$); the verbal channel (as opposed to mixed) was marginally significant ($\beta = -0.46, p = .07$), similar to the exemplar-based channel ($\beta = -0.44, p = .11$), all other effects were not significant. Overall, when controlling for teachers accuracy, we only see marginal beneficial effects of using mixed communication (as opposed to isolated channels), and the strong negative effect of a two-dimensional rule condition.

Lastly, it must be noted that interaction effects could not be reliably tested on the obtained data. Specifically, adding interactions between communication type and intervention variables results in unstable models, where conclusions highly depend on which interactions are included, while the natural choice of including all interactions of interest results in multicollinearity issues.

B.3.5. Teachers' subjective estimates of student performance

Teachers generally had a good grasp on how well their students were going to perform. Thus, the Kendall correlation between teacher's predictions about student performance and students actual accuracy is highly significant: $\tau = 0.352, p < .001$.

The estimate remains high for partial correlation controlling for teachers' accuracy ($\tau = .296, p < .001$). This shows that the correlation is not driven simply by teacher's awareness of their own knowledge, rather teachers are cognizant of difficulties of communicating knowledge in different conditions and/or are meta-cognitively aware about how good their teaching skills are relative to other participants.

B.4. Pilot experiment summary

First, teachers' communication volume depended on condition (category structure): teachers in more difficult conditions provided more information. Nevertheless, this adjustment was not sufficient, in the sense that all conditions still affected student performance (except for stimulus dimensionality which affects communication volume, but does not significantly affect student accuracy). Second, teachers adjusted not only the volume, but also the content of their messages in systematic ways, reflecting the difference in the structure of communicated categories. Third, mixed communication format resulted in higher student performance (significantly for verbal compared to mixed, marginally for exemplar-based compared to mixed). At the same time, teachers generally provided more information in the mixed condition. Thus, although mixed communication format was the most effective, it was not the most efficient among the three (per unit volume). Lastly, teachers demonstrated high awareness of the quality of their teaching materials. Specifically, teachers' estimates of their students' performance were significantly correlated with actual students' accuracies, even when teachers' mastery of category knowledge was controlled for.

Overall, the pilot experiment demonstrated the flexible nature of pedagogical category communication. Teachers, aware of the difficulties their students are facing, used a variety of techniques to adapt their messages to the category structure, changing both the volume and the content of their messages. Despite those adaptations, however, using a mixture of two different modes of communication was more effective than relying on isolated channels.

At the same time, we observed no specificity in how different channels are affected by confusability, dimensionality, or rule type. That is, our hypotheses stating that (a) verbal communication will be more robust to changes in stimuli dimensionality (b) exemplar communication will be more robust to confusibility, received no confirmation.

Although the results mentioned above are important on their own, the key role of the pilot experiment was to inform the hypotheses tested in Experiment 1.

Appendix C. Statistical analysis detail

In this section, we provide detailed specifications of the statistical models that we used for data analysis and give more detail for our overall approach in data analysis.

C.1. Main analyses with accuracy as target

As mentioned in our pre-registration (link), our main model in Experiments 1 and in the pilot was a binomial regression (a generalized linear model with a binomial random component and logistic link function), with inference done using robust ("sandwich") variance estimation. We used the R language **glm** function to fit the coefficients, and we used custom-written code for the robust variance estimation-based inference.[9] We provide our analysis code along with the submission. For main effects, unless noted otherwise, dimensionality was coded as a linear predictor, while confusability and channel were dummy-coded. We used "high" confusability and "mixed" communication as the base level for these variables. So overall, unless otherwise noted, the predictors were: confusability (dummy-coded), dimensionality (linear predictor with values 2, 3, or 4), communication channel (dummy-coded), teacher accuracy (logit-transformed). The latter is added to additionally control for teacher's performance, since we are interested in communication, not in potential differences in teacher's skill. This "teacher accuracy" predictor was never significant, however.

For models with interaction, we used an effect-coded version of the model above. Specifically, confusability and dimensionality were center-coded (+0.5, −0.5 for high and low values respectively). For communication channel, the following orthogonal-centered system with two variables was used: hyp1=(−1/3 when "channel" is either verbal or exemplar, and 2/3 for the mixed channel), hyp2=(1/2 for verbal, 0 for mixed, and −1/2 for exemplar).

We differed from our pre-registration in that we focused on predicting accuracy rather than accuracy gain per communication unit, since we realized the latter to be an extremely noisy measure.

The main model was chosen as the most robust, but not necessarily the most sensitive/powerful, since it does not capitalize on potential inter-dependencies between different data points (one teacher provided instructions for approximately three students). In our pre-registration (link), we mentioned that we planned to additionally use a generalized estimating equations model. However, our attempts to fit it in the first experiment resulted either in convergence issues or in unrealistically

[9] The code was written by Prof. Daniel Gillen and was provided as part of a graduate course on statistical inference taught in the University of California, Irvine. We provide it along with our analysis code.

optimistic estimates where almost all predictors became significant. We attributed it to the relatively low number of students per teacher, and relatively high noise, and decided to sacrifice potentially higher power of that model to ensure higher robustness and reliability of our results.

All main results and coefficients reported in the paper come from the binomial model with robust variance estimation. However, to ensure that the results we reported were not a consequence of a specific model choice, we ran two additional models in each statistical test: a beta-binomial model (using the **betabin** function from the **aod** R package), and, as a simple baseline and a sanity check, a standard binomial glm with naive (non-robust) inference (default inference in the R statistical package).

As expected, our model of choice (binomial glm with robust variance estimation) was indeed the most conservative. Also, as expected, vanilla (non-robust) glm shifted a few non-significant results into the region of significance (showing how dangerous is the common practice of using it as a default). Beta-binomial model coincides in its predictions with the robust binomial glm almost everywhere, although adding interactions in Experiment 1 did not change the significance of any of the coefficients into marginal significance (as it happened with the robust binomial model, see 2.2.1).

C.2. Additional analyses with integer targets

When the target was not accuracy, but an integer variable (as was the case specifically with predicting the number of words or exemplars in the Communication Volume section Section 2.2.2), we at first attempted to use a GLM with a log link and a Gaussian random component as our main model, with robust variance estimation. However, for the exemplar channel, we observed extreme differences between robust and non-robust estimation approaches, as well as high sensitivity to the choice of outlier treatment strategy (clipping vs removal) and threshold.

Therefore, we switched to simple and more robust rank-based methods instead. Specifically, we used Wilcoxon's rank-sum test to test the effect of confusability on volume, and Kendall's correlation to analyze the relationship between dimensionality and volume (since dimensionality has three levels, Wilcoxon's rank-sum test was not applicable).

C.3. Overall strategy

Overall, when the best model choice or variable coding was not obvious, we erred on the side of caution, running the analysis both ways, so that if the results are sensitive to the model choice or are otherwise unstable, we can report so in the paper.

Appendix D. Bayesian model detail

Since the distribution of students' accuracies in Experiment 2 (within specific conditions) was bimodal, with one peak at about 0.5 and the second peak higher, we needed to account for that when analyzing the data. The binomial glm that we used in Experiment 1 and in the pilot is unable to take advantage of such a structure, and can only capture the most general trends. A likely explanation for such a distribution is that a student either succeeds in understanding the gist of the communicated message and gets into the high-performing subgroup group, or fails to understand anything and performs at chance. A Bayesian mixture model is a natural choice for statistical analysis of such data.

We modeled student performance in each condition as a mixture of two distributions: the high-performing subgroup and the communication failure subgroup (performing at chance). Thus, every condition had two variables associated with it: (1) Probability of successful communication, denoted c. (2) Accuracy in the successful subgroup, i.e. the probability of giving a correct answer in the case of successful communication, denoted a.

To write the model formally, we are going to use the upper index to indicate communication channel, the first lower index to specify dimensionality (high or low), and the second lower index to specify confusability (high or low). For example, a^v_{hl} denotes the accuracy in the successful subgroup in the case of verbal communication with high dimensionality and low confusability. To estimate the overall effect of a given independent variable,[10] we look at the total difference between conditions corresponding to different levels of that variable. For example, for the verbal channel, the effect of dimensionality on the accuracy of the successful subgroup is measured as $(a^v_{hh}-a^v_{lh})+(a^v_{hl}-a^v_{ll})$.

The unsuccessful subgroup accuracy was fixed to 0.5 in all conditions, the successful subgroup accuracy prior was uniform between 0.5 and 1 for all conditions, and the probability of learning prior was uniform between 0 and 1.

The model was implemented in JAGS. The credible intervals reported in the paper are based on 10000 MCMC iterations, with 4 chains and a 10000 burn-in period.

Accounting for teacher accuracy. As formulated above, the model does not allow to control for teacher accuracy. We, therefore, also ran an alternative model, in which the "accuracy in the successful subgroup" coefficient is changed to "communication proportion in the successful subgroup", sampled uniformly from 0.6 to 1. The accuracy for a student in is then determined as this coefficient times the accuracy of the student's teacher. The value of 0.6 is chosen so that under minimal possible teacher accuracy (0.85), the resulting student accuracy is at least 0.5. Otherwise, the subgroup interpretations would start to overlap in highly undesirable ways.

We observed no qualitative differences in our results when this model was applied. In the paper, we reported the numbers obtained with the direct accuracy model since it is more standard and is easier to interpret.

Appendix E. Time-on-task reanalysis

During the review process, a number of concerns were raised regarding the possibility that the data might be polluted by overly fast students, who click through the experiment without engaging. To make sure that such students do not skew our results, we have eliminated them from the data and repeated our analyses.

E.1. Reanalysis procedure

There are different metrics of time-on-task that we could use to filter out suspiciously fast students. We opted to rely on "the time spent studying teaching materials", as opposed to the "median trial time" (a natural alternative). The reasoning behind this choice was as follows: if communication fails, i.e. if a student failed to understand the message from their teacher, it seems perfectly reasonable for them to click through the test examples. In other words, even participants who acted in good faith and made a fair effort to understand their teacher's message might "click through" the test stimuli if, despite their best effort, they failed to understand the message and do not know what to do.

With the key variable selected, we used the following exclusion strategy: in each experiment, we first identified the "minimal reasonable time", i.e. the minimal time that, at least for some students, was enough to succeed in the experiment. To do so, we found the fastest among all "successful" students, where "success" was defined as achieving an accuracy of 65% or above. The number 65 was initially chosen since with 30 binary trials, the probability of achieving such accuracy by chance is less than 5%, which seemed sufficiently low. By

[10] Here, in Experiment 1, we grouped dimensionalities 3 and 4 into one "high dimensionality" group, for simplicity and to keep the number of estimated coefficients down.

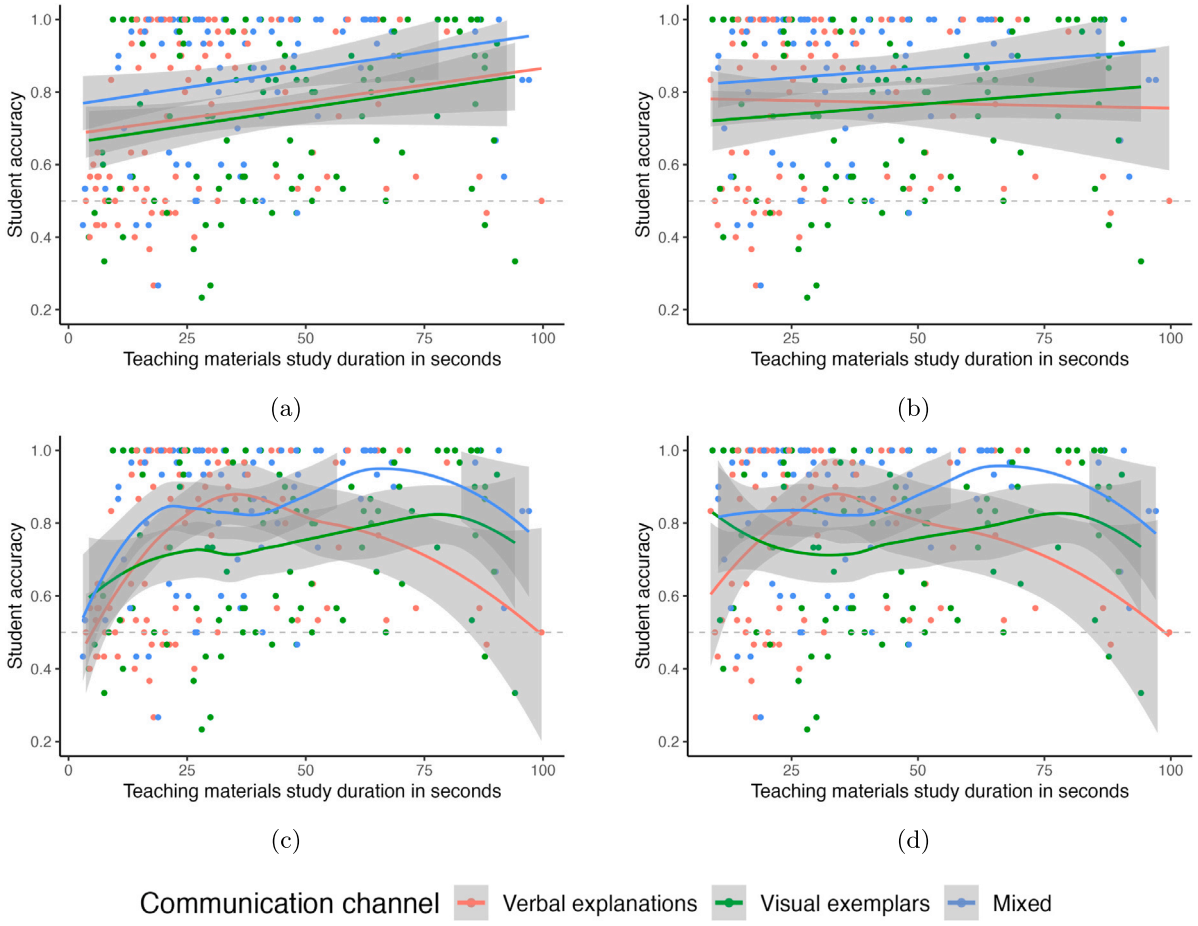

Fig. 8. Relationship between Study Duration and Student Accuracy for Different Communication Channels (study duration values above 100 were removed to aid visualization). Trend lines represent estimated linear relationships on the visualized data, and shaded areas indicate the 95% *CI*s. The top row corresponds to linear trend estimates, bottom row — LOWESS. Left column (a, c) - full data, right column (b, d) - anomalously fast students removed. The apparent upward trend is only present if the anomalously fast students are not removed from the visualization. Note that keeping or removing this group does not substantially affect any of our main results or conclusions (see Appendix E.).

Table 6
Summary statistics for time spent by students on studying teaching materials under different modes of communication in Experiment 1.

Communication channel	q10	q25	median	q75	q90	mean
Examples	13.51	28.5	45.61	70.9	89.68	73.76
Verbal	6.48	15.37	23.84	44.25	72.88	47.7
Mixed	14.29	23.06	40.65	77.12	186.24	71.41

coincidence, the two fastest successful students in both experiments happened to have an accuracy of 83.33, so the exact value of the "success" threshold ended up not being important, i.e. our reanalysis would yield exactly the same results with any "success" threshold under 83.33.

In Experiment 1, the fastest successful student studied their teaching materials 8.985 s, and in Experiment 2, the number was 5.382. This discrepancy may be due to the fact that teaching materials were generally much shorter and faster to study in Experiment 2, because of the volume restrictions.

Lastly, we removed all students who spent less than 8.985 studying teaching materials in Experiment 1, and students who spent less than 5.382 studying teaching materials in Experiment 2. With this filtered dataset, we repeated our previous analyses.

E.1.1. Reanalysis results

In Experiment 1, out of 316 students previously included in the analysis, we excluded 27, which still left a large number of "unsuccessful" students (84).

The median study time for these remaining unsuccessful students was 32.147, which again shows that failures to communicate did not all cluster near abnormally low times.

For the second experiment, only three out of 291 participants were removed.

When we repeated our previous analysis with this filtered data, most results remained qualitatively unchanged in direction and significance, with the following exceptions (all related to Bayesian analyses):

Changes in Bayesian analyses:

- Exp. 1: the negative influence of confusability on the probability of learning in the verbal channel changed from significant to marginally significant. New 95% credible interval: (−0.755, 0.065), old: (−0.819, −0.047).
- Exp. 2: the negative influence of dimensionality on the probability of learning in the example channel shifted from marginal significance to significance (new 95% credible interval: (−0.647, −0.107), old: (−0.611, 0.013)).

Table 7
Experiment 1: Regressing student accuracy on experimental conditions, controlling for teacher's accuracy (values without such a control are given in parentheses). The overall model is significant (deviance = 167.81, $df = 5, p_{\chi^2} < .001$).

	β	z-value	p-value
Intercept**	1.39 (1.34)	3.48 (4.25)	<.001 (<.001)
Low confusability**	0.46 (0.45)	2.69 (3.02)	.007 (.003)
Dimensions	0.01 (0.01)	0.14 (0.12)	.889 (.897)
Channel: exemplar (vs mixed baseline)**	−0.48 (−0.48)	−2.73 (−2.72)	.006 (.006)
Channel: verbal (vs mixed baseline)**	−0.55 (−0.55)	−3.09 (−3.09)	.002 (.002)
Logit of teacher accuracy	−0.02 (–)	−0.18 (–)	.854 (–)

Table 8
Experiment 1: Using effect-coded model, regressing student accuracy on experimental conditions, controlling for teacher's accuracy (values without such control are given in parentheses). The overall model is significant (deviance = 197.2, $df = 7, p_{\chi^2} < .001$).

	β	z-value	p-value
Intercept**	1.33 (1.26)	3.95 (16.8)	<.001 (<.001)
Confusability**	−0.48 (0.47)	−3.04 (3.02)	.005 (.002)
Dimensions	0.01 (0.01)	0.11 (0.09)	.916 (.925)
Channel h1 (Mixed vs Rest)**	0.52 (0.52)	3.01 (3.01)	.003 (.003)
Channel h2 (Verbal vs Exemplar)	−0.02 (−0.02)	−0.13 (−0.13)	.9 (.9)
Logit of teacher accuracy	−0.02 (–)	−0.19 (–)	.85 (–)
Interaction: h1 × Confusability	−0.11 (−0.11)	−0.33 (−0.33)	.747 (−0.742)
Interaction: h2 × Confusability	−0.62 (−0.62)	−1.86 (−1.86)	.063 (.063)

- Exp. 2: the negative influence of confusability on the probability of learning in the example channel shifted from marginal significance to non-significance (new 95% credible interval: (−0.555, 0.078), old: (−0.584, 0.041)).

Overall, there were no dramatic changes in any of the key effects. The advantage of mixed communication was completely unaffected. Of the less important effects mentioned in the discussion, two were affected. One shifted from marginal significance to significance, another — the other way around.

The overall changes in the second experiment, if anything, made its results more strongly aligned with our pre-registered hypotheses, since we expected example-based communication to be resilient against confusability and to be susceptible to changes in dimensionality.

That being said, we believe it is important not to overstate the consequences of these shifts in significance, especially for marginal effects, which are generally likely to wane in and out of significance based even on random sample perturbations.

Overall, none of the effects dramatically changed in direction, magnitude, or significance.

Appendix F. Time vs volume as the efficiency metric

In our experiments, we focused on studying efficiency as accuracy achieved by students per communication volume unit (i.e. per word or per example). It is important to mention that other efficiency metrics are possible, with "accuracy per unit time" being one natural alternative. Although our experiment was not designed to address the question of which channel was more efficient per unit time, we have observed some patterns that are worth mentioning as they might guide further investigation.

Table 6 presents the median study time for each condition in the experiment. The data indicate that the median study time in the verbal condition was considerably lower than in the mixed and exemplar conditions. Although it is possible that this reflects faster information comprehension via the verbal channel, we find it likely that this only reflects the user interface differences across conditions (see Appendix A). In the exemplar condition, participants were encouraged to perform an additional action by clicking on an exemplar to enlarge it and study it in detail, which likely contributed to a longer study time compared to the verbal condition.

At the same time, apart from the faster study time in the verbal condition, the general time-on-task patterns largely accord with the observations and conclusions we reached when analyzing communication volume (Section 2.2.2). As was mentioned in Appendix E, there is a small anomalously fast and poorly performing group (which, if excluded, does not substantially affect our results). Apart from it, as can be seen on Fig. 8, similarly to communication volume results, there is no general "more is better" effect when it comes to teaching materials study time (instead, each channel seems to have a "sweet spot" resulting in the best performance). Also, notably, mixed communication performs better across a wide range of study times.

Appendix G. Additional experimental results

Please see Table 7 to see additional results for Experiment 1 (coefficient estimates) for student performance regression. See Table 8 for the effect-coded interaction model results.

Appendix H. Content analysis

It is important to note that, due to the subjective nature of identifying the categories based on manual inspection, and since labeling was done by authors, rather than independent experts, the results in this section should be treated as illustrative.

H.1. Method

For all experiments, we analyzed the content of verbal messages created by teachers, classifying it into a number of typical types of messages. These types were identified based on the materials collected during the pilot study.

Specifically, through manual inspection, we identified seven common types of communicated information. For example, the "Exemplars" message type included messages that verbally describe members or prototypes of the categories being transferred, while the "Dimensionality reduction" message type included messages that explicitly indicate that certain dimensions are irrelevant (see Table 9 for definitions and examples of all message types). We evaluated all teachers' messages, identifying which message types they contain. Judgments were made by two authors independently solely based on teachers' texts. No other information was available during the evaluation to avoid possible biases. All disagreements were later resolved through discussion on a case-by-case basis.

Table 9
Types of teachers' verbal instructions and illustrative examples.

Instruction type & definition		Examples
Exemplars		
	Listing specific feature values that fit the category	"Type A fish have no tail. Type B have tails.", "Type A fish have their mouths either closed or slightly open, Type B fish have their mouths open wide."
Relative Rule		
	Values of the target attribute relative to another category	"Type A have shorter top fins compared to type B", "Type A tend to have darker colored undersides"
Dimensionality Reduction		
	Explicit indication of relevant or irrelevant dimensions	"Look at the color on the bottom", "Ignore everything on the fish except for the mouth"
Distribution		
	Optional information about the distribution of the exemplars along relevant or irrelevant dimensions	"There are ones with the spike on the head and then others without the spike", "The belly color of fish type A is always black", "tend to have", "usually has", "all have"
Boundaries and Threshold		
	Upper and lower boundaries of the category or a value that separates categories along the key dimension	"All else equal, look at the fins! Medium to long length is type B, short to short-medium is type A", "The cutoff between A and B is about midway between a triangle and a square shaped tail"
Strategies		
	Personal experience, heuristics, and metacognitive strategies useful for the task	"When in doubt, if the top fin looks like a triangle rather than a little stub, then it is likely type B", "It is the easiest way to tell between the two fish", "you will need to pay attention to how far open their mouths are"
Other		
	Reminding instructions, introducing definitions, and providing other information to the students.	"It is your goal to distinguish between two types of fish: A and B", "dorsal fin (the topmost fin on the fish's back)"

Table 10
Frequency of different message types in teachers' texts in all experiments.

	Pilot experiment	Experiment 1	Experiment 2
Exemplars	74%	73%	69%
Dimensionality Reduction	39%	44%	4%
Relative rule	17%	27%	27%
Distribution	12%	18%	11%
Boundaries and Threshold	18%	14%	11%
Strategies	19%	24%	1%
Other	7%	7%	4%

The overall goal of this analysis was to illustrate the content of the messages in a systematic way and to assess the relationship between stimulus characteristics (dimensionality and confusability) and the frequency of occurrence of different types of communicated information. Labeling of messages was not exclusive, as each teacher-generated verbal instruction could fall under more than one message type[11]. After all messages were labeled, we looked at distributions of these types of messages across conditions.

H.2. Content analysis results: Experiment 1

We performed content analysis on teacher-generated texts to see whether teachers adjusted the content of their messages in systematic ways depending on the condition they were put in. In our pilot study (see Appendix B), we identified seven common message types present in teachers' messages and then tagged all texts according to these types (each message may belong to more than one message type). See H.1 for detail on the approach. Most messages (73%) included descriptions of typical members of each category ("Exemplars" message type). Dimensionality Reduction (explicitly stating that certain stimuli dimensions are not informative) was the second most common type at 44%. All other message types were found in less than 30% of cases each (see Table 10).

Among the seven message types, we identified two, that, we expected, would be used more or less depending on our interventions. Specifically, we expected that higher stimulus dimensionality would result in a greater proportion of Dimensionality Reduction messages, while higher confusability would be associated with increased usage of Boundaries and Thresholds (to assist in distinguishing between borderline exemplars).

Teachers were indeed less likely to use Boundaries and Threshold messages in the low confusability condition (3% of cases) than in the high confusability one (24%), $\chi^2(1, N = 218) = 18.29, p < .001$. However, the effect of dimensionality on the frequency of Dimensionality Reduction messages between two- (39%), three- (55%), and four-dimensional (38%) conditions was only marginally significant ($\chi^2(2, N = 218) = 5.39, p = .067$). Given that it was also not monotone, we believe that it should be interpreted with caution. These results closely replicate those obtained in our pilot study (the only difference — in the pilot, the Dimensionality Reduction result was not significant, rather than marginally significant).

The key takeaway of this analysis is that the contents of verbal messages that teachers generate differ in systematic ways, depending on condition.

Message type distribution across all experiments is given in Table 10. One can see that in the restricted volume condition (Experiment 2) some key message types such as "Exemplars" and "Relative rule"

[11] For example, consider "The difference between type A fish and type B fish is the size of the dorsal fin (the topmost fin on the fish's back). Type A fish have a very short and small dorsal fin whereas type B fish have a much taller dorsal fin. Mouth position (how open or closed it is) and rear fin size can both vary between type A and type B fish and therefore are not useful in distinguishing between the two". The "Type A fish have a very short and small dorsal fin" snippet belongs to the "Exemplars" category, while "type B fish have a much taller dorsal fin" is a relative rule.

remain as frequent as in the first two experiments. At the same time, "Dimensionality reduction" and "Strategies" message types experience a precipitous drop. This shows that, although highly frequent in Experiments 1 and 2, these types are secondary and are often sacrificed when communication volume is restricted.

Category usefulness. To see whether any of the identified categories were particularly helpful (or, on the contrary, harmful) in category communication, we ran an additional exploratory analysis. Specifically, we added the indicator variables for each of the strategies to the main regression predicting student accuracy. As before we used a binomial glm with robust variance estimation. Other predictors we controlled for were communication channel (in this case only mixed or verbal, since the categories are not applicable to the exemplar channel), confusability, dimensionality, and the logit of teacher accuracy. None of the predictors were, however, significant.

H.3. Content analysis results: Experiment 2

The content of teachers' verbal messages mostly contained descriptions of typical Exemplars and Relative Rules, as before. However, the proportion of Dimensionality Reduction messages (4%) dropped substantially compared to Experiment 1 (44%). A similar decline was found for the use of Strategies (1% compared to 24% in Experiment 2). We attribute this to the word limits that were introduced in this experiment. As in the first experiment, high confusability increased the use of Boundaries and Thresholds from 3 to 20%. This time, however, the effect was only marginally significant ($\chi^2(1, N = 71) = 3.68, p < .055$). No statistically significant differences in Dimensionality Reduction were found between two- (6%) and four-dimensional (3%) conditions ($\chi^2(1, N = 71) < 0.01, p = .980$).

References

Aboody, R., Velez-Ginorio, J., Laurie, R., Santos, L. R., & Jara-Ettinger, J. (2018). When teaching breaks down: Teachers rationally select what information to share, but misrepresent learners' hypothesis spaces. In *Proceedings of the 40th annual meeting of the cognitive science society, vol. 1* (pp. 72–77).

Anderson, J. R. (1990). *The adaptive character of thought*. Hillsdale, NJ, US: Lawrence Erlbaum Associates, Inc.

Anderson, J. R. (1991). The adaptive nature of human categorization. *Psychological Review*, *98*(3), 409.

Aodha, O. M., Su, S., Chen, Y., Perona, P., & Yue, Y. (2018). Teaching categories to human learners with visual explanations. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 3820–3828).

Ashby, F. G., Alfonso-Reese, L. A., Waldron, E. M., et al. (1998). A neuropsychological theory of multiple systems in category learning. *Psychological Review*, *105*(3), 442.

Ashby, F. G., & Maddox, W. T. (2005). Human category learning. *Annual Review of Psychology*, *56*(1), 149–178. http://dx.doi.org/10.1146/annurev.psych.56.091103.070217.

Ashby, F. G., Paul, E. J., & Maddox, W. T. (2011). COVIS. In E. M. Pothos, & A. J. Wills (Eds.), *Formal approaches in categorization* (pp. 65–87). Cambridge University Press, http://dx.doi.org/10.1017/CBO9780511921322.004.

Avrahami, J., Kareev, Y., Bogot, Y., Caspi, R., Dunaevsky, S., & Lerner, S. (1997). Teaching by examples: Implications for the process of category acquisition. *The Quarterly Journal of Experimental Psychology Section A*, *50*(3), 586–606. http://dx.doi.org/10.1080/713755719.

Bandura, A. (1977). *Prentice-Hall series in social learning theory, Social learning theory*. Englewood Cliffs, New Jersey: Prentice-Hall.

Bridgers, S., Jara-Ettinger, J., & Gweon, H. (2020). Young children consider the expected utility of others' learning to decide what to teach. *Nature Human Behaviour*, *4*(2), 144–152.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., et al. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, *33*, 1877–1901.

Chopra, S., Tessler, M. H., & Goodman, N. D. (2019). The first crank of the cultural ratchet: Learning and transmitting concepts through language. In *Proceedings of the 41st annual meeting of the cognitive science society* (pp. 226–232).

Clark, A. (1998). Magic words: How language augments human computation. In *Language and thought: Interdisciplinary themes* (pp. 162–183). Cambridge University Press.

Cohen, A. L., Nosofsky, R. M., & Zaki, S. R. (2001). Category variability, exemplar similarity, and perceptual classification. *Memory & Cognition*, *29*(8), 1165–1175.

Dubova, M., & Goldstone, R. L. (2021). The influences of category learning on perceptual reconstructions. *Cognitive Science*, *45*(5), Article e12981.

Gentner, D. (2016). Language as cognitive tool kit: How language supports relational thought. *American Psychologist*, *71*(8), 650.

Hutchins, D., Schlag, I., Wu, Y., Dyer, E., & Neyshabur, B. (2022). Block-recurrent transformers. In A. H. Oh, A. Agarwal, D. Belgrave, & K. Cho (Eds.), *Advances in neural information processing systems*. Retrieved from https://openreview.net/forum?id=uloenYmLCAo.

Keren, G., & Schul, Y. (2009). Two is not always better than one: A critical evaluation of two-system theories. *Perspectives on Psychological Science*, *4*(6), 533–550.

Kloos, H., & Sloutsky, V. M. (2008). What's behind different kinds of kinds: effects of statistical density on learning and representation of categories. *Journal of Experimental Psychology: General*, *137*(1), 52.

Kotov, A. A., & Kotova, T. N. (2018). The role of different types of labels in learning statistically dense and statistically sparse categories. *The Russian Journal of Cognitive Science*, *5*(3), 56–67, Retrieved 2020-01-05, from http://www.ssrn.com/abstract=2580829.

Li, T. J.-J., Radensky, M., Jia, J., Singarajah, K., Mitchell, T. M., & Myers, B. A. (2019). Pumice: A multi-modal agent that learns concepts and conditionals from natural language and demonstrations. In *Proceedings of the 32nd annual ACM symposium on user interface software and technology* (pp. 577–589).

Lupyan, G. (2012). What do words do? Toward a theory of language-augmented thought. In *Psychology of learning and motivation, vol. 57* (pp. 255–297). Elsevier.

Maddox, W. T., & Ashby, F. G. (2004). Dissociating explicit and procedural-learning based systems of perceptual category learning. *Behavioural Processes*, *66*(3), 309–332.

Malte, A., & Ratadiya, P. (2019). Evolution of transfer learning in natural language processing. arXiv preprint arXiv:1910.07370.

Minda, J. P., & Miles, S. J. (2010). The influence of verbal and nonverbal processing on category learning. In B. H. Ross (Ed.), *Psychology of learning and motivation, vol. 52* (pp. 117–162). Academic Press, http://dx.doi.org/10.1016/S0079-7421(10)52003-6.

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.

Moskvichev, A., & Liu, J. A. (2021). Updater-extractor architecture for inductive world state representations. arXiv preprint arXiv:2104.05500.

Moskvichev, A., Tikhonov, R., & Steyvers, M. (2019). A picture is worth 7.17 words: Learning categories from examples and definitions. In *Proceedings of the 41st annual meeting of the cognitive science society* (pp. 2406–2412). Montreal, Canada: Cognitive Science Society, Retrieved from https://mindmodeling.org/cogsci2019/papers/0416/0416.pdf.

Newell, B. R., Dunn, J. C., & Kalish, M. (2011). Systems of category learning. In *Psychology of learning and motivation, vol. 54* (pp. 167–215). Elsevier, http://dx.doi.org/10.1016/B978-0-12-385527-5.00006-1.

Nosofsky, R. M., Sanders, C. A., Gerdom, A., Douglas, B. J., & McDaniel, M. A. (2017). On learning natural-science categories that violate the family-resemblance principle. *Psychological Science*, *28*(1), 104–114.

Rips, L. J. (1989). Similarity, typicality, and categorization. In S. Vosniadou, & A. Ortony (Eds.), *Similarity and analogical reasoning* (1st ed.). (pp. 21–59). Cambridge University Press, http://dx.doi.org/10.1017/CBO9780511529863.004.

Rosedahl, L., & Ashby, F. G. (2018). A new stimulus set for cognitive research. http://dx.doi.org/10.31234/osf.io/2xz3q.

Rosedahl, L., Serota, R., & Ashby, F. G. (2021). When instructions don't help: Knowing the optimal strategy facilitates rule-based but not information-integration category learning. *Journal of Experimental Psychology: Human Perception and Performance*, *47*(9), 1226–1236. http://dx.doi.org/10.1037/xhp0000940.

Seger, C. A., & Miller, E. K. (2010). Category learning in the brain. *Annual Review of Neuroscience*, *33*, 203–219.

Shafto, P., Goodman, N. D., & Griffiths, T. L. (2014). A rational account of pedagogical reasoning: Teaching by, and learning from, examples. *Cognitive Psychology*, *71*, 55–89.

Shepard, R. N., Hovland, C. I., & Jenkins, H. M. (1961). Learning and memorization of classifications. *Psychological Monographs: General and Applied*, *75*(13), 1.

Sloutsky, V. M. (2010). From perceptual categories to concepts: What develops? *Cognitive Science*, *34*(7), 1244–1286.

Sloutsky, V. M., et al. (2016). Selective attention, diffused attention, and the development of categorization. *Cognitive Psychology*, *91*, 24–62.

Smith, E. E., & Sloman, S. A. (1994). Similarity-versus rule-based categorization. *Memory & Cognition*, *22*(4), 377–386. http://dx.doi.org/10.3758/BF03200864.

Sumers, T. R., Ho, M. K., Hawkins, R. D., & Griffiths, T. L. (2023). Show or tell? Exploring when (and why) teaching with language outperforms demonstration. *Cognition*, *232*, Article 105326. http://dx.doi.org/10.1016/j.cognition.2022.105326.

Tomasello, M. (1999). The human adaptation for culture. *Annual Review of Anthropology*, *28*(1), 509–529. http://dx.doi.org/10.1146/annurev.anthro.28.1.509.

Vong, W. K., Hendrickson, A. T., Navarro, D. J., & Perfors, A. (2019). Do additional features help or hurt category learning? The curse of dimensionality in human learners. *Cognitive Science*, *43*(3), Article e12724.

Vygotsky, L. (1978). *Mind in society: The development of higher psychological processes*. Cambridge: Harvard University Press.

Yu, A., & Mooney, R. J. (2022). Using both demonstrations and language instructions to efficiently learn robotic tasks. arXiv preprint arXiv:2210.04476.

Zettersten, M., & Lupyan, G. (2018). Using language to discover categories: More nameable features improve category learning. In *Proceedings of the 12th international conference on the evolution of language*. Wydawnictwo Naukowe Uniwersytetu Mikołaja Kopernika, http://dx.doi.org/10.12775/3991-1.136.

Zettersten, M., & Lupyan, G. (2020). Finding categories through words: More nameable features improve category learning. *Cognition*, *196*, Article 104135. http://dx.doi.org/10.1016/j.cognition.2019.104135.