Short communication

Show or tell? Exploring when (and why) teaching with language outperforms demonstration

Theodore R. Sumers a,*, Mark K. Ho a, Robert D. Hawkins b, Thomas L. Griffiths a,b

a Department of Computer Science, Princeton University, Princeton, NJ, United States of America
b Department of Psychology, Princeton University, Princeton, NJ, United States of America

ARTICLE INFO

Keywords:
Pedagogy
Communication
Language
Demonstration
Abstraction

ABSTRACT

People use a wide range of communicative acts across different modalities, from concrete demonstrations to abstract language. While these modalities are typically studied independently, we take a comparative approach and ask when and why one modality might outperform another. We present a series of real-time, multiplayer experiments asking participants to teach concepts using either demonstrations or language. Our first experiment (N = 416) asks when language might outperform demonstration. We manipulate the complexity of the concept being taught and find that language communicates complex concepts more effectively than demonstration. We then ask why language succeeds in this setting. We hypothesized that language allowed teachers to reference abstract object features (e.g., shapes and colors), while demonstration teachers could only provide concrete examples (specific positive or negative objects). To test this hypothesis, our second experiment (N = 568) ablated object features from the teacher’s interface. This manipulation severely impaired linguistic (but not demonstrative) teaching. Our findings suggest that language communicates complex concepts by directly transmitting abstract rules. In contrast, demonstrations transmit examples, requiring the learner to infer the rules.

1. Introduction

Human teaching takes many forms. For instance, imagine teaching a friend how to play chess and trying to convey how rooks move. You could demonstrate the rule by dragging a rook along a row, then a column. Alternatively, you could describe the rule with language, saying “Rooks move along rows and columns” without ever touching a piece. These distinct modalities both play a role in human pedagogy—but what are their strengths, and how do they work? When and why is language more effective than demonstration?

Researchers have taken different approaches to this question. Classic work has provided rich qualitative insights into natural settings (Carroll & Bandura, 1990; Chi, 2013; Scribner & Cole, 1973). For example, younger children tend to teach via nonverbal demonstration, while older children and adults tend to use language (Ellis & Rogoff, 1982; Strauss, Ziv, & Stein, 2002). Meanwhile, more tightly controlled experiments have tested quantitative predictions from computational theories (e.g., Shafto, Goodman, & Griffiths, 2014), providing mechanistic insight into specific behaviors such as demonstration (Buchsbaum, Gopnik, Griffiths, & Shafto, 2011; Gweon, Tenenbaum, & Schulz, 2010; Ho, Cushman, Littman, & Austerweil, 2021), verbal explanation (Chopra, Tessler, & Goodman, 2019), or evaluative feedback (Ho, Cushman, Littman, & Austerweil, 2019). Yet this work has largely focused on individual teaching modalities without considering relationships between them (but see Tessler, Bridgers, & Tenenbaum, 2020). This paper seeks a middle ground, introducing a novel interactive teaching paradigm to directly compare modalities while controlling for factors that may mediate their efficacy.

We report a pair of studies that compare two communication modalities—demonstration and language. Demonstration (showing someone how chess pieces move) is a form of example-based teaching (Bandura, 1965; Csibra & Gergely, 2009; Shneidman, Gweon, Schulz, & Woodward, 2016). Demonstrations are non-symbolic communicative acts, relying on shared understanding of a problem domain (such as physical affordances and goals; Gergely & Csibra, 2003; Gergely & Jacob, 2012). The capacity to generate and understand demonstrations is foundational, emerging early in childhood (Butler & Markman, 2012; Gweon & Schulz, 2018; Gweon et al., 2010; Király, Csibra, & Gergely, 2013; Leonard, Lee, & Schulz, 2017). Language (e.g., telling someone the rules of chess) is another cognitively privileged mechanism for transmitting knowledge (Grice, 1975). Language relies on shared conventions (Clark, 1996) to convey information about categories (Gelman & Markman, 1986, 1987), relations (Loewenstein & Gentner,
Concepts can be more or less complex. Intuitively, the rule for how rooks move is simpler than the rule for casting. Increasing complexity makes individual learning more difficult (Feldman, 2000; Shepard et al., 1961) and favors abstract problem representations (Koedinger, Alibali, & Nathan, 2008). We thus predicted that both modalities would suffice for teaching simple concepts, but language could better convey abstract information, while demonstrations instantiate that information in concrete examples (Csibra & Shamsudheen, 2015).

But when and why might one be preferred? We develop a pair of concept-teaching studies to address these questions. Experiment 1 asks when language outperforms demonstration, varying the teaching context by manipulating the concept complexity (Feldman, 2000; Shepard, Hovland, & Jenkins, 1961). We find that concept complexity has systematic and differential effects on these modalities: language outperforms demonstration when communicating complex concepts. We then ask why, hypothesizing that language succeeds by encoding the abstract, underlying rule (Lombrozo, 2006). Experiment 2 tests this hypothesis by varying the teaching context. We ablate the teacher’s view of object features, while leaving spatial context intact, effectively forcing teachers to communicate about specific objects rather than abstract object features. We find that this manipulation severely impairs linguistic teaching while leaving demonstrative teaching unaffected. This systematic comparison supports the idea that language relies on shared abstractions to efficiently transmit complex concepts. In contrast, demonstrations struggle to communicate such concepts, but are less reliant on shared abstractions.

2. Experiment 1: Linguistic teaching conveys complex concepts better than demonstrations when shared abstractions are available

Concepts can be more or less complex. Intuitively, the rule for how rooks move is simpler than the rule for casting. Increasing complexity makes individual learning more difficult (Feldman, 2000; Shepard et al., 1961) and favors abstract problem representations (Koedinger, Alibali, & Nathan, 2008). We thus predicted that both modalities would suffice for teaching simple concepts, but language could better convey complex concepts. To test this prediction, we introduce a collaborative teaching game that places pairs of participants in an interactive virtual environment. The environment contains colored shapes with different point values. One participant, the teacher, can see the objects’ values but is unable to act in the environment. Consequently, they must communicate this information to their partner, the learner, who controls an avatar in the environment. We manipulated the communication modality and concept complexity in a fully between-subjects design. In the language condition, teachers sent chat messages to the learner; in the demonstration condition, teachers produced movement trajectories instead (see Fig. 1). Finally, we manipulated complexity by changing the mapping of point values to object visual features, giving rise to different underlying concepts.

2.1. Methods

2.1.1. Participants
We recruited 480 participants using Prolific (www.prolific.co). Participants were required to be fluent in English, possess an approval rating of 95%, and be located within the United States, United Kingdom, or Ireland. They were paid $2.25 with a score-based bonus up to $1.75. We excluded seven participants who accidentally re-started the experiment after refreshing their browser, for a final sample of 227 pairs, 208 of whom completed all 10 levels.

2.1.2. Procedure
After reading instructions, all participants completed two practice trials to familiarize them with the environment. Participants were then paired, assigned teacher or learner roles, and played a practice trial together. They then played 10 trials of a communication task. Each trial proceeded as follows. First, in the action phase (Fig. 1A), the learner was given 30 s to study the level and plan their movements. They then had 8 s to move the avatar to collect objects while the teacher watched. The learner was only shown the shapes and colors, but the teacher could also see the underlying values. After 8 s, the action phase ended and both players were shown the net score. They then proceeded to a teaching phase (Fig. 1B). Language teachers were allowed to message their partner through a one-way chat interface; demonstration teachers re-played the level while the learner watched. After 10 trials, both players completed an exit survey, which included a measure of concept knowledge: learners were asked the value of each of the 8 possible objects (+1/-1/"Don’t Know").

1 “A king can be moved two spaces towards a rook, and the rook to the opposite side of the king, if the king is not in check, neither piece has previously moved, there are no pieces between them, and the king does not pass through a square that would place it in check.”

2 The study was approved by the Princeton IRB; our sample size and planned analyses were pre-registered at https://aspredicted.org/s85_fmq. Code and data are available at https://github.com/tsumers/show-or-tell.

3 For example, collecting three objects with a value of +1 and one with a value of −1 would yield a net score of 2.
2.2.1. Transmission fidelity

Our primary hypothesis concerned transmission fidelity. We expected that linguistic teachers could convey the underlying concept more effectively than demonstrative teachers, especially for more complex concepts. We operationalized transmission fidelity using learner concept accuracy, the number of correct responses in the exit survey (Fig. 2A). To test our hypothesis, we used a binomial logistic regression predicting the number of correct responses (out of 8). We included predictors for communication modality, (Boolean) concept complexity, and their interaction. To interpret main effects, we centered the regression model using three levels of complexity — low, intermediate, and high, corresponding to Shepard concepts I, II, and III (Shepard et al., 1961) or Boolean complexity 1, 4, and 6, in the taxonomy of Feldman (2000). Within each complexity condition, we randomized features to mitigate salience biases. To allow direct comparison of movement trajectories, values of objects in each location were generated randomly but fixed across participants for each trial. Thus all participants proceeded through the same set of 10 randomly generated maps, but with different visual features assigned to locations.

2.2.2. Concept simplification

Our analysis of transmission fidelity asked how much concept knowledge learners acquired. We now investigate what they learned. Rather than make random errors, we hypothesized that demonstration learners would formulate a simpler rule based on fewer features (Goodman, Tenenbaum, Feldman, & Griffiths, 2008), thereby reducing its complexity. For example, a demonstration learner assigned the most complex concept in Fig. 1C (Boolean complexity 6) might observe their teacher collect mostly triangles and infer that all and only triangles were positive (Boolean complexity 1). To operationalize this prediction, we compared the Boolean complexity of the learner’s survey responses to the true concept (Fig. 2B). We ran a regression analysis predicting the complexity of the learner’s inferred concept as a function of the communication modality (contrast-coded as before), controlling for true complexity (dummy-coded with low as the baseline). We found a weak effect of modality ($\beta = .24$, $t(202) = 2.01$, $p = .045$; see Table S4 for full model). While all teachers struggled to convey the most complex concepts, demonstration learners simplified even intermediate ones.

2.3. Discussion

As predicted, linguistic teaching outperformed demonstration at transmitting complex concepts. Demonstration learners tended to prefer simple explanatory rules; in contrast, language allowed teachers to more precisely specify complex feature combinations. Of course, factors aside from modality may inhibit learning of more complex concepts. For example, teachers may have deliberately simplified harder concepts, teaching “good enough” rules intended to maximize the learner score rather than transmit the full concept (see Table S6 for an example of teaching a simplified concept first). Learning may also be affected by more general factors such as memory and attention. However, because these factors were held constant across conditions, our results reflect modality-based differences for the communicative pair as a whole.

These findings suggest that when teachers know the abstract features that define a concept, language matches or outperforms demonstration. But how does language transmit such complex concepts? We hypothesized that language succeeds by allowing direct reference to these task-relevant abstractions (Loewenstein & Gentner, 2005; Ratterman & Gentner, 1998). Our second experiment ablates these features, where we predict that language—but not demonstrations—would be severely impaired.

---

4 For example, in the low complexity condition, a single dimension determined the value. Pairs could be assigned blue as positive (pink as negative), pink as positive (blue as negative), squares as positive (triangles as negative), and so on.

5 We treated complexity as a continuous variable and subtracted the mean, $(1 + 4 + 6)/3$. We sum coded the communication condition (language = 1, demonstration = −1).
3. Experiment 2: Linguistic (but not demonstrative) teaching relies on shared abstractions

Experiment 1 found that language outperforms demonstration at communicating complex concepts. Why might this be the case? Consider the statement “Rooks move along rows and columns.” Rook is an abstraction: it refers to a set of objects defined by the label rook (Gelman, 2004; Gelman & Markman, 1986). But what if one did not have such a label? One could imagine referring to specific rooks: “That piece in the corner moves along rows and columns.” This reference creates ambiguity when it comes time for the listener to generalize. Which other pieces does this same property extend to? What about the opponent’s identically-shaped (but differently colored) pieces? Intuitively, such reasoning seems cumbersome. Indeed, in Experiment 1, most language teachers expressed rules by referring to abstract object features: “filled pink shapes and hollow blue shapes” (Fig. 1B), rather than specific objects. In contrast, demonstrations cannot refer to abstract features: physical actions necessarily engage with physical objects.

This observation motivates our hypothesis: such shared abstractions underpin language’s success in Experiment 1. To test this hypothesis, we introduced a new manipulation: we removed the abstract object features (shape, color, and fill) from the teacher’s interface. Teachers in the “Locations and Features” condition had the same interface as Experiment 1. Locations only teachers were given an ablated interface: they saw the value of objects, but not their features (shape, color, and fill; Fig. 3A). Otherwise, the procedure was the same as Experiment 1, using the same sequence of 10 trials.

3.2. Results

3.2.1. Transmission fidelity

As in Experiment 1, our primary variable of interest was the learner’s concept knowledge (Fig. 4A). We hypothesized locations only context would negatively affect linguistic teaching, without impairing demonstration. To test for an interaction between communication modality and context, we again used a binomial logistic regression predicting correct responses on the learner’s exit survey. Predictors included complexity, communication modality, context (all effect-coded), and the communication-context interaction. All four terms were significant, critically including the communication-context interaction ($\beta = .34, t(279) = 6.90, p < .001$; see Table S9), consistent with our hypothesis that locations only context would impair linguistic teaching more than demonstration (see Appendix B for full regression models).

3.2.2. Cumulative scores

Task performance may provide another perspective on communicative success: did different teaching modalities allow learners to reach higher scores? Specifically, we examine the cumulative score (Fig. 4B). If two learners end the experiment with equal concept knowledge but different scores, this implies that the higher-scoring learner acquired the concept faster. We used a linear regression to predict each learner’s cumulative score using same predictors as above. The results suggest even more extreme limitations of locations only context linguistic teaching. Even in the low complexity condition, these pairs scored worse than

---

Fig. 3. Manipulating access to abstractions in Experiment 2. A: Under the locations only condition, teachers could not see object features. They could only see which objects were positive. B/C: Locations only affected linguistic teaching, but not demonstrations. Top: demonstration trajectories from the second level. Teachers in both conditions relied on score-maximizing demonstrations; here, they overwhelmingly collected the five positive objects in the bottom right, then moved as far as possible towards the top right, which was the next best corner. Bottom: most common words in each condition. Linguistic teachers in the two conditions used very different strategies: Locations and features teachers relied primarily on the features (shape, color, and fill), while Locations only teachers relied on behavioral (“first 3 positive”) or spatial (“top left the top 2 are negative”) references to describe which objects were positive.

---

6 Our pre-registration is available at https://aspredicted.org/ZHS_YW1.
pairs in either demonstration condition (interaction $\beta = 4.13, t(279) = 5.58, \ p < .001$, see Fig. S4 and Table S10 for full model).

3.2.3. Teacher strategies and subjective difficulty

How did removing object features affect teaching? Language teachers in different context conditions used very different words (Fig. 3B/C; Appendix C). Features and locations teachers relied heavily on the abstract features (color, shape, and fill), while locations only teachers primarily referenced specific positive objects (e.g., using spatial references).

In contrast, demonstration teachers were largely unaffected by context. Teachers’ most common strategy was to maximize their score: 75% (1078 of 1430) of demonstrations achieved the highest possible points per level. There was not a significant difference in teachers’ rates of score-maximizing demonstrations across the four conditions ($F(3, 139) = 1.515, p = .213$). In other words, most demonstration teachers taught by acting optimally—a viable strategy even when the teacher and learner have differing representations of the environment.

Finally, we explored the division of labor across conditions and roles by asking participants to rate how difficult the game was (Fig. 5). Participants generally found the low complexity condition easy. However, at intermediate complexity, theoretically interesting differences emerged across conditions and roles ($F(7, 278) = 18.76, p < .0001$; see Table S21 for Tukey HSD tests). Language teachers and learners agreed that sharing features and locations was easier than locations only. Both roles thought features and locations was easy; both thought locations only was hard. In contrast, in demonstration conditions, learners found the task more difficult than teachers; the shared context did not matter. This asymmetry suggests that demonstrative teaching places the cognitive burden on the learner: it was easy for teachers to provide demonstrations, but difficult for learners to infer the underlying concept.

3.3. Discussion

Our findings support the hypothesis that language succeeded in Experiment 1 by relying on shared abstractions: allowing reference to object features rather than the objects themselves. Lacking access to these features, linguistic teaching suffered while demonstrations did not. We note that even with locations only, linguistic teachers were in principle capable of communicating more information than demonstrative teachers. Indeed, one locations only language teacher specified the value of every object in every trial (Table S18). In general, however, such strategies were difficult and inefficient: despite sending more text (Fig. S2), locations only teachers struggled to transmit the concept.8

4. General discussion

Teaching is central to human social life and has been examined extensively from evolutionary (Tomasello, 2009), developmental (Csibra & Gergely, 2009), cross-cultural (Shneidman et al., 2016) and computational (Shafto et al., 2014) perspectives. Yet recent experimental work has largely studied single modalities without considering their relative efficacy. Our study compared different modalities in a rich but controlled communication task.

In our experiments, language teachers excelled at communicating abstract information. However, without access to object features, they found it challenging to communicate the same information in terms of specific objects. In contrast, demonstration teachers struggled to communicate complex concepts, but were largely unaffected by removing object features. This pattern highlights a key difference between modalities: language may use abstract labels to convey abstract information, whereas demonstrations use concrete examples, forcing the learner to infer abstract information (Butler & Markman, 2016; Hume, 1748). While humans possess a remarkable ability to learn from such data (Baldwin, Markman, & Melartin, 1993; Butler & Markman, 2012; Csibra & Gergely, 2009; Gergely, Egyed, & Király, 2007; Shafto et al., 2014), a conventionalized lexicon allows language to directly encode abstract information (Csibra & Shamanskeen, 2015; Gelman & Markman, 1986). Notably, while learning from expert demonstrations is a historically popular approach in robotics (Abbeel & Ng, 2004; Argall, Chernova, Veloso, & Browning, 2009), recent trends in the field emphasize learning from language instead (Luketina et al., 2019; Tellex, Gopalan, Kress-Gazit, & Matuszek, 2020).

However, the present study captures only a small slice of human communication. We enforced one-way communication, hindering discourse repairs (Dingemanse et al., 2015) and ad-hoc formation of new linguistic abstractions (McCarthy, Hawkins, Wang, Holdaway, & Fan, 2021) or signaling conventions (Galantucci & Garrod,

---

8 We note two possible concerns regarding locations only linguistic teaching. First, the condition may have been difficult for teachers to understand. Examination of chat logs revealed several teachers who either failed to understand how the learner’s perspective differed from theirs or offered encouragement rather than teaching. Removing these pairs, however, did not affect our results (Appendix D). Second, the lack of spatial structure made referencing specific objects particularly challenging. However, a followup study with more spatial structure replicated our findings (Appendix F).
In addition, we studied transmission of a particular kind of information: Boolean concepts (Shepard et al., 1961). Tradeoffs between modalities may differ when teaching other types of concepts (e.g., fundamentally physical athletic concepts, Martens, 1975). Indeed, despite extensive specialized vocabularies, demonstrative teaching is widespread in domains such as music (Dickey, 1991; Weeks, 1996), dance (Keenlyside, 2010), cooking (Mondada, 2014), sports (Evans & Reynolds, 2016), martial arts (Ramam, 2019), medical procedures (Svensson, Luff, & Heath, 2009) and vocational training (Asplund & Kilbrink, 2018). Finally, we studied modalities independently; naturalistic teaching may interleave them, such as narrating a demonstration (Carroll & Bandura, 1990) to establish shared linguistic abstractions.

Our findings suggest several avenues of future work. First, our experimental paradigm may support extending computational models to interactive multi-modal teaching. Communication is classically viewed as a means to intervene on abstract beliefs (Goodman & Frank, 2016; Grice, 1975; Shafto, Wang, & Wang, 2021), but it is inherently grounded in real-world actions: humans necessarily infer partners' beliefs through observations (Baker, Jara-Ettinger, Saxe, & Tenenbaum, 2017) and integrate information from different modalities (Butler & Tomasello, 2016; Carroll & Bandura, 1990; Davis-Unger & Carlson, 2008). Thus, models of communication situated within decision-making environments (Bridgers, Jara-Ettinger, & Gweon, 2019; Fisac et al., 2020; Sumers, Hawkins, Ho, & Griffiths, 2021) could begin to explain interactive, multi-modal communicative acts.

Finally, transmission fidelity is central to accounts of cumulative cultural evolution (Lewis & Laland, 2012; Tomasello, 2009), but the question of which modalities are necessary or sufficient to accumulate knowledge remains open. Empirical work has used tasks such as stone tool making (Lombao, Guardiola, & Mosquera, 2017; Morgan, Uomini, Rendell, Chouinard-Thuly, Street, Lewis, Cross, Evans, Kearney, de la Torre, Whiten, & Laland, 2015) or complex problem solving (Dean, Kendal, Schapiro, Thierry, & Laland, 2012; Tessler, Tsvidis, Madeano, Harper, & Tenenbaum, 2021; Thompson, van Opheusden, Sumers, & Griffiths, 2022), but measuring knowledge accumulation in these domains is challenging. In contrast, Boolean concepts allow us to precisely track what individuals learned. Future work could extend our paradigm to iterated settings to directly measure the acquisition and refinement of complex concepts over multiple generations.

**CRediT authorship contribution statement**

**Theodore R. Sumers**: Conceptualization, Software, Formal analysis, Investigation, Data curation, Visualization, Funding acquisition, Writing. **Mark K. Ho**: Conceptualization, Formal analysis, Supervision, Writing. **Robert D. Hawkins**: Conceptualization, Formal analysis, Supervision, Writing. **Thomas L. Griffiths**: Conceptualization, Funding acquisition, 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.

**Acknowledgments**

The authors declare that there is no conflict of interest. TRS is supported by the NDSEG Fellowship Program and RDH is supported by the National Science Foundation (grant #1911835). This work was additionally supported by the National Science Foundation (grant number #1544924) and the John Templeton Foundation (grant #61454). We thank M.H. Tessler and our anonymous reviewers for insightful and constructive feedback.

**Appendix A. Supplementary data**

Supplementary material related to this article can be found online at [https://doi.org/10.1016/j.cognition.2022.105326](https://doi.org/10.1016/j.cognition.2022.105326).
T.R. Sumers et al.

Cognition 232 (2023) 105326


