Psychon Bull Rev (2015) 22:673–686
DOI 10.3758/s13423-014-0734-y

THEORETICAL REVIEW

Learning multisensory representations for auditory-visual transfer of sequence category knowledge: a probabilistic language of thought approach

Ilker Yildirim · Robert A. Jacobs

Received: 15 January 2014 / Revised: 29 August 2014 / Accepted: 4 September 2014 / Published online: 23 October 2014

Abstract If a person is trained to recognize or categorize objects or events using one sensory modality, the person can often recognize or categorize those same (or similar) objects and events via a novel modality. This phenomenon is an instance of cross-modal transfer of knowledge. Here, we study the Multisensory Hypothesis which states that people extract the intrinsic, modality-independent properties of objects and events, and represent these properties in multisensory representations. These representations underlie cross-modal transfer of knowledge. We conducted an experiment evaluating whether people transfer sequence category knowledge across auditory and visual domains. Our experimental data clearly indicate that we do. We also developed a computational model accounting for our experimental results. Consistent with the probabilistic language of thought approach to cognitive modeling, our model formalizes multisensory representations as symbolic "computer programs" and uses Bayesian inference to learn these representations. Because the model demonstrates how the acquisition and use of amodal, multisensory representations can underlie cross-modal transfer of knowledge, and because the model accounts for subjects' experimental performances, our work lends credence to the Multisensory Hypothesis. Overall, our work suggests that people automatically extract and represent objects' and events' intrinsic properties, and use these properties to process and understand the same (and similar) objects and events when they are perceived through novel sensory modalities.

Keywords Multisensory perception · Language of thought · Sequence learning · Computational modeling

Introduction

Human cognition is robust, at least in part, because people mentally represent objects and events in a variety of ways, such as perceptual, motoric, and semantic representations. Even within perception, people represent objects and events in multiple ways. This fact is demonstrated by cross-modal transfer of knowledge. If a person is trained to visually categorize a set of objects, this person will often be able to categorize novel objects from the same categories when objects are grasped but not seen (Wallraven, Bülthoff, Waterkamp, van Dam, & Gaißert, 2014; Yildirim & Jacobs, 2013). Because knowledge acquired during visual learning is used during haptic testing, this finding suggests the existence of both visual and haptic representations of objects. Below, we report an experiment in which people were trained to either auditorily or visually categorize sequences of events. When tested with sequences presented in a novel sensory modality, people were often able to categorize these sequences too. Because training

We thank S. Piantadosi for many helpful discussions, and the editor and anonymous reviewers for constructive comments on this manuscript. This work was supported by research grants from the National Science Foundation (DRL-0817250, BCS-1400784) and the Air Force Office of Scientific Research (FA9550-12-1-0303).

I. Yildirim (✉) · R. A. Jacobs
Brain and Cognitive Sciences, Massachusetts Institute of Technology, 77 Massachusetts Ave, Cambridge, MA 02139, USA
e-mail: ilkery@mit.edu

I. Yildirim · R. A. Jacobs
Laboratory of Neural Systems, The Rockefeller University, 1230 York Ave, New York, NY, 10065, USA

R. A. Jacobs
Brain and Cognitive Sciences, University of Rochester, Rochester, NY, 14627, USA
e-mail: robbie@bcs.rochester.edu

and testing used different modalities, this result indicates that people had representations of event sequences in both modalities.

How do people transfer knowledge across sensory modalities? A plausible hypothesis, referred to here as the Multisensory Hypothesis, is that people use sensory-specific representations of objects and events to infer amodal or multisensory[1] representations characterizing objects' and events' intrinsic properties. These representations facilitate cross-modal transfer of knowledge. To understand this hypothesis, it is important to recognize the distinction between objects' and events' intrinsic (or "deep") properties and the sensory (or "surface") features that these properties give rise to. For instance, the location of an event is a modality-independent intrinsic property. Visual and auditory features are modality-dependent sensory cues to the event's location arising when the event is viewed or heard, respectively. To explain how the Multisensory Hypothesis accounts for cross-modal transfer, consider a sequence categorization task. For example, sequences of events moving in a clockwise direction belong to category *A*, whereas sequences moving in a counterclockwise direction belong to category *B*. When a person is trained to visually categorize event sequences, the person uses his or her visual representations to infer multisensory representations characterizing sequences' intrinsic properties. When subsequently tested with an auditory sequence, the person judges its category based on whether it is more consistent with the intrinsic properties of sequences belonging to category *A* or category *B*.

The Multisensory Hypothesis predicts that people acquire modality-independent representations of objects' and events' intrinsic properties. Converging neural, behavioral, and computational evidence suggests that this is the case. A striking example comes from Quiroga (2012) who argued that human brains contain "concept" cells which are involved in the representation of individual people or objects regardless of the modality used to sense those people or objects. For instance, when recording in the human medial temporal lobe, he and his colleagues reported a neuron that selectively responded when a person viewed images of the television host Oprah Winfrey, viewed her written name, or heard her spoken name (Quiroga, Kraskov, Koch, & Fried, 2009). These and similar findings indicate that our brains encode abstract representations that are amodal or multisensory in the sense that they are activated by perceptual inputs spanning multiple modalities.

Why focus on the Multisensory Hypothesis? The Multisensory Hypothesis is an appropriate focus because of its inherent interest and potential importance. Due to their abstract, modality-independent nature, multisensory representations are a form of conceptual representation. Currently, the field of Cognitive Science knows very little about how people acquire conceptual representations from sensory data, though this topic has garnered much interest in recent years (e.g., see the literature on grounded cognition; Barsalou, 2008). Furthermore, the study of multisensory perception is attracting much attention (Calvert, Spence, & Stein 2004, Stein, 2012). It may be that recent advances in our understanding of multisensory perception shed new light on the Multisensory Hypothesis.

To our knowledge, no one has attempted to explicitly define and implement a model of cross-modal transfer based on the Multisensory Hypothesis. The sole exception is our earlier work where we showed how multisensory representations of object shape—consisting of representations of an object's parts and the spatial relations among these parts—can be acquired from visual or haptic features, and showed how these representations can facilitate transfer of object category knowledge across visual and haptic modalities (Yildirim & Jacobs, 2013. The work reported in this article builds on our earlier work. However, it studies a new domain, namely cross-modal transfer of sequence category knowledge across visual and auditory modalities. In addition, it uses a different modeling approach.

In cognitive modeling, one school of thought favors symbolic approaches, such as approaches based on production rules or logic. Another school of thought favors statistical approaches, such as approaches based on connectionist networks or Bayesian inference. Advocates of these different schools of thought have different perspectives, and have often engaged in heated debates (McClelland & Patterson, 2002a, b; Pinker & Ullman, 2002a, b). Unfortunately, these debates have not led to a resolution as to which framework is best.

Our viewpoint is that both symbolic and statistical frameworks have important merits, and thus it may be best to pursue a hybrid approach taking advantage of each framework's best aspects. This viewpoint is recently emerging in the Cognitive Science literature (e.g., Goodman, Tenenbaum, Feldman, & Griffiths, 2008; Piantadosi, Tenenbaum, & Goodman, 2012; Ullman, Goodman, & Tenenbaum, 2012). It is referred to as a "probabilistic language of thought" (pLOT) approach because it applies Bayesian inference to a representation consisting of symbolic primitives and combinatorial rules (Fodor 1975). To date, the pLOT approach has been used almost exclusively in domains that are typically modeled using symbolic methods, such as human language and high-level cognition. A contribution of the work presented here is that we apply this approach to the study of

[1] Terms such as 'multisensory', 'amodal', 'modality-independent', and 'modality-invariant' often have slightly different meanings to different people. Consequently, here, we often ignore these reader-dependent and subtle differences and use these terms in an interchangeable manner. We believe that this will not lead to confusions so long as the terms are interpreted in context.

human perception, an area whose study is dominated by statistical techniques. We believe that the pLOT approach can be advantageous for characterizing perceptual processes, particularly multisensory processes, including the acquisition of amodal, multisensory representations of objects and events from sensory data and their subsequent use.

Experiment

Previous experimental and theoretical studies examined people's performances in tasks requiring them to learn about sequences. For example, researchers studied the learnability of sequences with different kinds of structural (e.g., Markovian, non-Markovian, hierarchical) dependencies (e.g., Jordan, 1986; Elman, 1990; Cleeremans & McClelland, 1991; McCallum, 1996; Fiser & Aslin, 2002), and proposed different kinds of cognitive architectures to explain the observed behavioral patterns (see Gureckis & Love, 2010, for a critical review and comparison).

A subset of these researchers used sequences of spatial locations (e.g., Hunt & Aslin, 2001; Deroost & Soetens, 2006; Hunt & Aslin, 2010; Bo & Seidler, 2010). The serial reaction time task is frequently used in these studies. It has been found that people's reaction times decline more quickly with a structured or highly predictable sequence than with a random or relatively unpredictable sequence (e.g., Hunt & Aslin, 2001).

Our experiment focuses on categorization of spatial sequences, and on generalization of sequence category knowledge to exemplars presented in an untrained sensory modality. The experiment made use of an innovative auditory-visual environment (see Fig. 1a) whose major components are a vertically-oriented (and oriented perpendicular to a subject's line of sight) planar surface covered with sheet metal, speakers, and light emitting diodes (LEDs). Each speaker and LED has a magnet attached to its back, meaning that each speaker and LED can be placed at any location on the vertical surface. Because a scrim (a curtain made from light gauzy material often used in theatre productions) covers the environment, the speakers and unlit LEDs are not visible by a subject. However, lit LEDs are visible to a subject due to the scrim's translucent properties. This environment is very useful for auditory-visual experiments. It is a large-scale environment—when a subject is seated 60 cm from the vertical surface, speakers and LEDs can be placed over a region subtending nearly 90 degrees of visual angle. The environment is flexible because speakers and LEDs can be placed at any location on the vertical surface, and precise because each speaker and LED is controlled independently on a millisecond time scale.

Participants

Participants were 21 students from the University of Rochester. All participants were at least 18 years old. We obtained all participants' written informed consent. Each experimental session lasted less than an hour, and participants were paid $10. This study was approved by the University of Rochester Research Subjects Review Board.

Stimuli

Stimuli consisted of temporal sequences of spatial locations presented in the auditory-visual environment. There were 7 possible locations arranged on an imaginary circle of radius about 57 cm (see Fig. 1b). Sequence lengths were sampled from a uniform distribution with minimum and maximum values of 6 and 15, respectively. When a sequence was presented auditorily, a location was indicated by a beep emitted by a small speaker. When a sequence was presented visually, a location was indicated by a flash of a white LED. Beeps or flashes lasted 200 ms, and pauses between beeps or flashes lasted 300 ms.

Sequences were exemplars from 4 possible categories. Fourteen exemplars from each category were generated. For

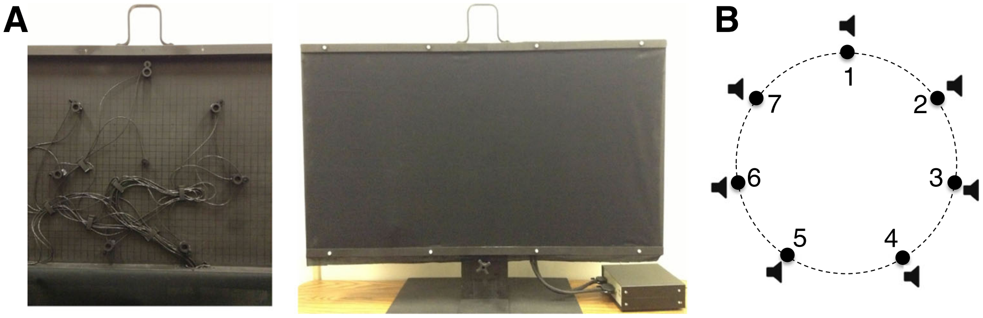


Fig. 1 **a** Photos of the audio-visual environment. In the left photo, the speakers, LEDs, and electrical hardware are visible. In the right photo, a scrim conceals the environment, meaning that the speakers, unlit LEDs, and electrical hardware are not visible. **b** A schematic of the 7 locations used in our experimental stimuli, and the speaker and LED at each location

each category, each of the 7 possible locations was used as a starting location twice.

Locations in exemplars from Category 1 change one unit in a clockwise direction at each time step, referred to as a [+1] pattern. Using the location indices in Fig. 1b, exemplars from Category 1 include “456712”, “2345671234”, and “45671234”.

Exemplars from Category 2 are clockwise cycles of length 3, denoted [+1 +1 -2]. That is, the second location is one clockwise unit from the first location, the third location is one clockwise unit from the second location, and the fourth location is equal to the first location. This pattern repeats until the end of the sequence. Exemplars include “23423423”, “7127127”, and “456456456”.

For Category 3, exemplars are counterclockwise cycles of length 4, denoted [-1 -1 -1 +3]. Exemplars include “17651765176517”, “6543654”, and “54325432”.

Exemplars in Category 4 follow a [+2 -1] pattern. Exemplars include “72132435465”, “132435”, and “4657617213”.

Procedures

At the start of a trial, a red LED at the center of the auditory-visual environment was illuminated for 1000 ms. Participants were asked to fixate this LED. Next, a sequence was presented, either auditorily or visually. Following the sequence presentation, participants indicated the category to which they thought the sequence belonged by pressing a key on a keyboard. On training trials, auditory feedback indicated whether a response was correct. Feedback was not provided on test trials.

Participants were seated approximately 50 cm from the auditory-visual display panel. However, because people’s auditory estimates of location are less accurate than their visual estimates (Battaglia, Jacobs, & Aslin, 2003; Alais & Burr, 2004), and because localization of auditory events was difficult during preliminary studies, participants were encouraged to lean forward to be as close as possible to the panel while observing auditory events.

Two groups of 9 people each participated in the experiment (3 people were excluded because they performed at chance on training trials or because they did not complete the full set of training and test trials). Participants in Group A-V were told at the start of the experiment that they would receive auditory training followed by visual testing. During training, these participants were trained to categorize 36 exemplars (9 exemplars from each of the 4 categories selected at random for each participant) presented auditorily. This auditory training stage consisted of blocks of 36 trials, where all exemplars were presented once and in randomized order in a block. At the end of a block, a message appeared on a computer screen informing a participant of his or her performance during that block. During training, a participant’s performance was monitored within a window of the most recent 36 trials. Training was terminated as soon as this performance exceeded 90 %, or when the participant completed 7 blocks of training.

Following training, Group A-V participants were reminded that test trials would use the visual modality. Test trials were identical to training trials except that sequences were presented visually, and participants did not receive feedback about the correctness of their responses. Participants performed 56 test trials (14 exemplars from each of 4 categories; 9 of the 14 exemplars were familiar [these sequences were presented during auditory training], whereas 6 exemplars were novel). Presentation order of the test sequences was randomized.

Participants in Group V-A followed the same procedures as participants in Group A-V except that the training and test modalities were switched. These participants underwent visual training and auditory testing.

Results

The left and right panels of Fig. 2 show each participant’s learning curve during training for Groups V-A and A-V, respectively. The horizontal axis of each graph plots the training trial number, and the vertical axis plots the percent correct in the most recent 36 trials (for trials up to the 36th, we assumed that a participant made $36 - t$ incorrect responses where t is the trial number). Participants in Group V-A were better at learning the categories (9 of 9 participants from Group V-A and 3 of 9 participants from Group A-V reached the training cut-off criteria of 90 %). The performances of Group A-V participants tended to plateau at around the 100th trial. These differences in the learning performances between the two groups are most likely due to the differing reliabilities of audition and vision for spatial localization. Clearly, however, all participants acquired significant knowledge of the sequence categories (chance performance is 25 %).

The left panel in Fig. 3 shows participants’ average performances on the final training block (i.e., the last 36 trials of training) and on the test block for both groups (error bars indicate standard errors of the means). The training performance of Group V-A reflects the fact that all participants achieved the training cut-off criteria of 90 % correct. On auditory test trials, the performance of this group remained high (about 75 %). The drop in performance from training to test is most likely due to the lower reliability of audition for spatial localization. The training performance of Group A-V was also good (slightly less than 75 %), and its test performance was not significantly different than its training performance ($t = -0.8$, $DF = 15.3$, $p = 0.44$,

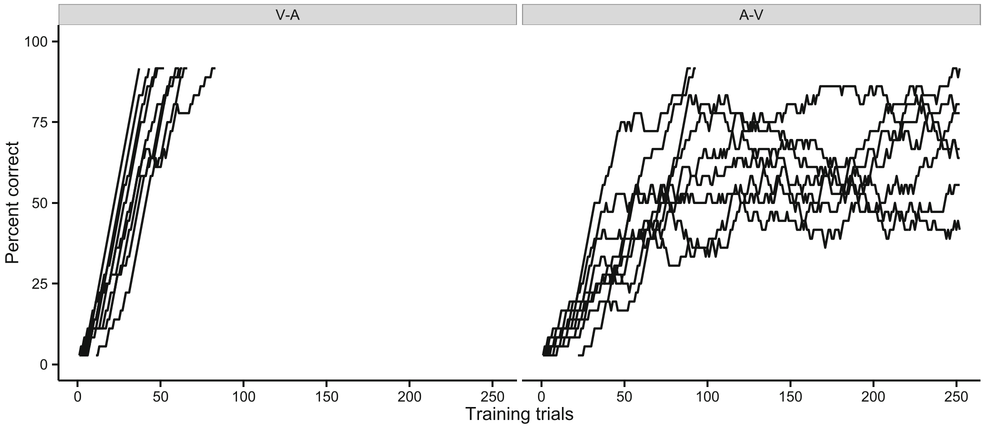


Fig. 2 Learning curves during training for Groups V-A and A-V

two-tailed t-test with unequal variances). Neither Group V-A's nor Group A-V's test performances differed on trials with familiar (i.e., previously observed during training) versus novel sequences (Group A-V: $t = 0.5$, $DF = 16$, $p = 0.61$; Group V-A: $t = -0.4$, $DF = 15.2$, $p = 0.67$; both two-tailed t-tests with unequal variances).

The right panel in Fig. 3 shows the final training block and test performances across the two groups when trials are sorted by the category of the sequence observed on a trial. For example, the leftmost bar in this panel shows the average performance on the final training block on trials that used sequences that are exemplars from Category 1. When examining the data in this manner, chance performance is 50 % correct because a participant either correctly classified a sequence from, for instance, Category 1 or did not. This analyses allows us to examine the relative ease of correctly classifying exemplars from each category. When sorted by the categories, a Friedman test revealed that there was a statistically significant rank ordering of the categories across training ($p < 0.001$) and test ($p < 0.01$) blocks. From participants' average performances, categories can be ordered with respect to their learnability (from highest to lowest) as follows: Category 1, Category 4, Category 2, and Category 3.

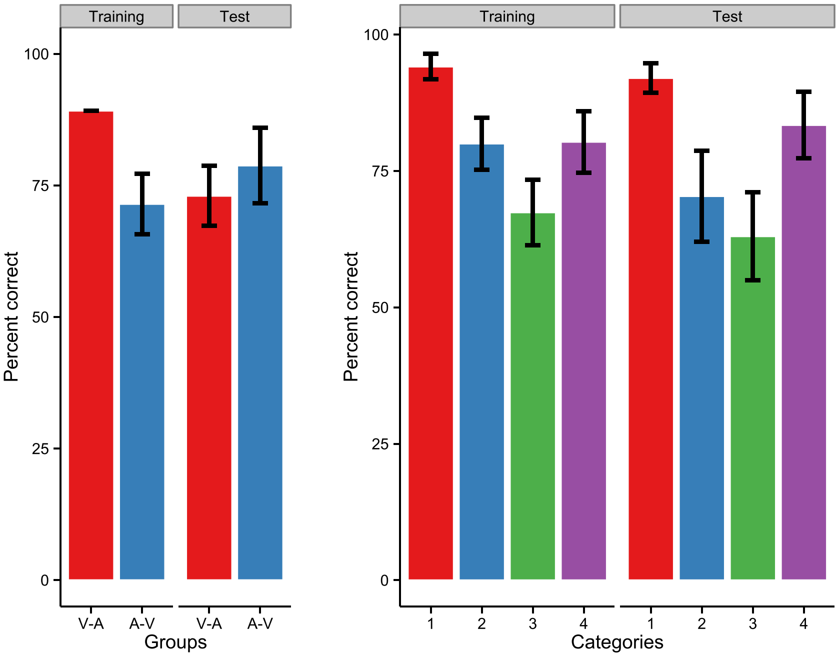


Fig. 3 (*Left*) Average performances on the final training block (last 36 trials of training) and on the test block for Groups V-A and A-V (error bars indicate standard errors of the means). (*Right*) Groups' final-training block and test performances when trials are sorted by the category of the sequence observed on a trial

In summary, we are interested in people's abilities to acquire and transfer knowledge of sequence categories. When categorical knowledge of spatial sequences is obtained through one sensory modality, can people transfer this knowledge to conditions in which sequences are observed through an untrained modality? Our experimental results indicate that the answer is yes.

Overview of the computational model

Our experiment suggests that participants transferred sequence category knowledge across auditory and visual modalities. How did they do this? To address this question, we propose a computational model accounting for our experimental results. The model includes a multisensory representation of each sequence category. A multisensory representation characterizes the intrinsic properties of a category in a modality-independent manner (Yildirim & Jacobs, 2013). The model also includes sensory-specific forward models. Because sensory-specific forward models map multisensory representations to sensory data, they can be thought of as implementing a type of mental imagery (Miall & Wolpert, 1996; Ito, 2008; Tian & Poeppel, 2010; Yildirim & Jacobs, 2013). This section provides an overview of the model's components. The next section describes learning and cross-modal transfer by the model in the context of our experiment.

Multisensory representations of sequence categories

Given auditory exemplars, visual exemplars, or both, the model learns a multisensory representation of a category. Behavioral and neural data suggest the existence of multisensory representations, and also suggest that these representations underlie, at least in part, a variety of behaviors in auditory-visual environments (e.g., Calvert et al. 1997; Pascual-Leone & Hamilton, 2001; Pekkola, et al., 2005; Tanabe, Honda, & Sadato, 2005; de Gelder & Vroomen, 2000; von Kriegstein & Giraud, 2006; Lehmann & Murray, 2005; Quiroga, Kraskov, Koch, & Fried, 2009; Liang, Mouraux, Hu, & Iannetti, 2013).

We characterize multisensory representations as "computer programs" for generating or predicting exemplars from a category. This approach builds on earlier work by Piantadosi et al. (2012) who used computer programs to characterize people's mental representations of numerical concepts.

When designing the computational model, our main focus was not on developing new insights regarding how people acquire and process sequential information. Although this is an important topic, many researchers already study this topic (e.g., Jordan, 1986; Elman, 1990; Cleeremans & McClelland, 1991; Gureckis & Love, 2010; McCallum, 1996; Fiser & Aslin, 2002). Rather, our primary goal was to understand how multisensory representations can be learned from sensory data, and to understand how multisensory representations can facilitate transfer of knowledge across sensory modalities. Because our model needs to represent sequences, it necessarily resembles previously existing models that also represent sequences. Of particular interest is the fact that our model shares important features with an early model of sequence learning by Simon and Kotovsky (1963). Although our model and their model have different goals, the two models use "programming languages" with similar symbolic operators to represent sequences. Indeed, it is only a moderate stretch to say that our model might be seen as a revised version of their model, modified to accept and process sensory data and updated to use modern learning (Bayesian inference) techniques.

A ***Ring***

```
L1: init(k)
L2: next(k)
L3: go to L2
```

B ***Length of 3***

```
L1: init(k)
L2: next(k)
L3: next(next(k))
```

C ***Forward, backward***

```
L1: init(k)
L2: next(k)
L3: prev(k)
L4: go to L2
```

D ***Loop***

```
L1: init(k)
L2: prev(k)
L3: prev(k)
L4: prev(k)
L5: go to L1
```

Fig. 4 Sample programs for characterizing sequence categories

We describe the programming language used by our model by explaining several sample programs. Consider the program in Panel A of Fig. 4. This program generates sequences in which locations change one unit in a clockwise direction at each time step (Category 1 from the experiment reported above). The first line of the program, denoted `L1`, randomly initializes a spatial cursor, denoted `k`. The spatial cursor is a variable that keeps track of the current spatial location. The `init` function randomly sets the cursor to a random integer between 1 and 7 (recall that there are 7 possible locations). Line `L2`, `next(k)`, moves the cursor one unit in a clockwise direction. Line `L3`, `go to L2`, states that the next line to be executed is `L2`, thereby creating a loop. Putting aside the fact that the program creates infinite sequences of locations (see below), the reader should intuitively understand that this program is consistent with sequences such as "456712" and "23456", but inconsistent with sequences such as "124" and "765".

Next, consider the program in Panel B. It uses the same primitives as the previous program, but it composes them in a different way. This program generates sequences of length 3 in which the second location is one clockwise unit from the first location, and the third location is two clockwise units from the second location (e.g., sequences such as "124" and "457"). In addition, this program uses recursion (see line `L3`).

The program in Panel C generates sequences that alternate between two neighboring locations, first moving one clockwise unit, then returning to the original location by moving one counterclockwise unit. Movement of the spatial cursor by one counterclockwise unit is achieved using the command `prev(k)` (line `L3`). This program is consistent with sequences such as "454545" and "1212121".

The program in Panel D generates sequences with counterclockwise cycles of length 4. It is consistent with sequences such as "654365" and "32173217". This program illustrates an additional feature of the `init` function that was not illustrated by earlier programs. The first time that `init` is called, it sets the spatial cursor to a random location. It then stores this location. Subsequent calls to `init` set the cursor to the stored location.

Based on these programs, the reader should have a good understanding of the nature of the model's programming language. Programs contain line numbers, a spatial cursor, and `init`, `next`, `prev`, and `go to` commands. Programs are capable of looping and of recursion. Clearly, these elements provide the model with a rich, expressive language for characterizing sequence categories.

Sensory-specific forward models

Because multisensory representations are modality-independent, sensory-specific forward models are needed to relate the representations to sensory data. An auditory-specific (vision-specific) forward model maps an exemplar to a prediction of the auditory (visual) features that an observer would perceive when the exemplar is auditorily (visually) rendered. Because of the simple nature of our stimuli—beeps and flashes—our forward models are relatively simple.[2] In particular, the location of an observed beep or flash is predicted to be equal to the location of the actual beep or flash plus some additive noise sampled from a circular Gaussian or von Mises distribution (recall that locations lie on a circle). The key feature of these forward models are their noise distributions. Because vision is a more precise cue to spatial location than audition (Battaglia et al. 2003; Alais & Burr, 2004), the vision-specific forward model used a noise distribution with a higher precision ($\kappa_V = 4.0$ roughly corresponding to a variance of 17°) than the noise distribution used by the audition-specific forward model ($\kappa_A = 2.5$ corresponding to a variance of 32°). The values of κ_A and κ_V were chosen on the basis of a trial-and-error search for values that allowed the model's predictions to match our experimental data. This occurred whenever vision was a reasonably more precise cue to spatial location than audition—that is, the model's performance was highly robust to the exact values chosen.

Learning and cross-modal transfer

Having introduced the multisensory representations and sensory-specific forward models, we now describe how the model learns and transfers knowledge across auditory and visual modalities. We do so in the context of the experiment described above.

The model learns multisensory representations of sequence categories based on its sensory input. As described above, the hypothesis space of category representations (i.e., the space of possible computer programs) is large. How should the model evaluate different hypotheses during learning? Here, we cast this problem as an instance of Bayesian inference.

For ease of exposition, we describe the model from the standpoint of a participant in Group V-A. Let $V = \{\vec{v}_1, \ldots, \vec{v}_N\}$ denote N visual sequences from one of the four categories (we exclude subscripts indexing categories to avoid unnecessary notation). Each $\vec{v}_i$ is a vector of K_i spatial locations, where K_i is the length of the i^{th} visual sequence. Thus, we write $\vec{v}_i = [v_{i1}, \ldots, v_{iK_i}]^T$, and let v_{ij} denote the visual observation at the j^{th} time step in sequence $\vec{v}_i$.

The model learns multisensory representations from sensory data as follows. Cognitive models often assume that sensory data are the products of a generative process. In the context of our experiment, a visual sequence is generated when a multisensory representation for a sequence category produces a sequence of locations and this sequence is visually rendered. To learn about multisensory representations, this generative process can be inverted via Bayes' rule:

$$P(R|V) \propto P(R)\,p(V|R) = P(R)\prod\nolimits_{i=1}^{N}\prod\nolimits_{j=1}^{K_i} p(v_{ij}|R) \quad (1)$$

where $P(R)$ is the prior probability of multisensory representation R, and $p(V|R)$ is the likelihood function arising from the vision-specific forward model. We consider each of these quantities—the prior and the likelihood function—in turn.

[2] In other cases, sensory-specific models can be complex. For example, Yildirim and Jacobs (2013) considered visual and haptic cues to object shape. In this case, the vision-specific forward model was a graphics library (e.g., OpenGL) and the haptics-specific forward model was a simulator of a human hand.

Table 1 PCFG for multisensory representations of sequence categories

Production rule		Probability

S →	`init(L)`
	`U`
		1.0

U →	`L`	0.25
U →	`O`	0.25
U →	`L` `U`	0.25
U →	`O` `U`	0.25
L →	`init(L)`	0.25
L →	`next(L)`	0.25
L →	`prev(L)`	0.25
L →	`k`	0.25
O →	`go to [one of the earlier lines in the current program (use equal probabilities)]`	1.0

As described above, multisensory representations are computer programs. For the purpose of assigning prior probabilities to programs, we characterize these programs using the probabilistic context-free grammar (PCFG) in Table 1 (this general approach is adopted from Piantadosi et al. (2012)).[3] A particular program, R, can be generated from the start symbol S by a derivation, a sequence of productions in the PCFG that ends when all non-terminals are replaced with terminals. At each step of a derivation, a choice is made among the productions which could be used to expand a non-terminal. Because a probability is assigned to each production choice in a derivation, the probability of the complete derivation is the product of the probabilities for these choices. In principle, the prior probability of a program should be defined as the sum of the probabilities of its possible derivations. However, derivations using our grammar are unique due to the structure of the grammar (there is, at most, only one non-terminal on the right-hand side of a production rule) and the consistent order in which we expand the non-terminals. In accord, we calculate the prior probability of program R, $P(R)$, using the equation:

$$P(R) = P(\mathcal{T}|\mathcal{G},\rho) = \prod\nolimits_{n\in\mathcal{N}_{nt}} P(n \to ch(n)|\mathcal{G},\rho) \tag{2}$$

where $\mathcal{T}$ is the derivation (i.e., parse tree) for program R, $\mathcal{G}$ is the set of production rules in the PCFG (left column of Table 1), and ρ is the set of probabilities associated with the production rules (right column of Table 1). In addition, N_{nt} is the set of all non-terminals in derivation $\mathcal{T}$, $ch(n)$ is the set of node n's children nodes, and $P(n \to ch(n)|\mathcal{G},\rho)$ is the probability for production rule $n \to ch(n)$.[4]

An advantage of this prior distribution is that it favors "simple" programs, meaning programs with short derivations. (To see this, note that Eq. 2 multiplies probabilities [i.e., numbers less than one]. The number of terms that are multiplied increases with the length of the derivation.) Consequently, it can be regarded as a type of Occam's Razor.[5]

The likelihood of a visual sequence, $p(\vec{v}_i|R)$, was estimated as follows. The initial observed location, v_{i1}, is an imperfect cue to the actual starting location of a sequence due to sensory noise. To deal with this uncertainty, we used the vision-specific forward model to select the three most probable locations based on the value of v_{i1}, and averaged the likelihood scores over these locations:

$$p(\vec{v}_i|R) = \frac{1}{3}\sum\nolimits_{l_1\in L} p(\vec{v}_i|R,l_1) \tag{3}$$

where $l_1 \in L$ indexes the three most probable locations, and $p(\vec{v}_i|R,l_1)$ is the likelihood score of sequence $\vec{v}_i$ based on multisensory representation R assuming that the sequence started at location l_1.[6]

To compute $p(\vec{v}_i|R,l_1)$, we used program R to generate a sequence. The initial location of this sequence was set to l_1. If the program was not capable of generating a sequence whose length is at least as long as K_i—the length of visual sequence $\vec{v}_i$—then the likelihood score was set to 0 (e.g., the program in Panel B of Fig. 4 only generates sequences of length 3). Otherwise the program was used to generate a sequence of length K_i. Let l_j denote the j^{th} element of this sequence. The likelihood score of $p(\vec{v}_i|R,l_1)$ is computed using the vision-specific forward model as follows:

$$p(\vec{v}_i|R,l_1) = \prod\nolimits_{j=1}^{K_i} VM(v_{ij}|l_j,\kappa_V) \tag{4}$$

[3] Although the expressions that our PCFG support are not Turing complete, we refer to these expressions as "computer programs" because, intuitively, the expressions resemble programs. In principle, the PCFG introduced here can be extended to support Turing complete computation.

[4] This prior probability distribution is similar (but not identical) to the prior distribution used by Goodman et al. (2008) which analytically integrates out ρ by assuming it has a uniform hyper-prior distribution. Our choice of prior distribution is motivated by the fact that Piantadosi et al. (2012) reports that choosing ρ to be uniform over the production rules for each non-terminal gives very similar results to integrating it out.

[5] *Ceteris paribus*, moves that require recursion (e.g., moving the location of the cursor two steps clockwise) are less probable under this distribution. Recursion is appealing from the perspective of computability. We believe that a deliberate experimental design can shed light on the trade-off between inserting new primitives to a grammar versus recursive calls.

[6] We also considered a different likelihood function in which, instead of summing over possible initial locations, we searched for the initial location that maximized the likelihood score of $\vec{v}_i$ with respect to R. Our simulation results were qualitatively indistinguishable between these two alternatives.

where $VM(\cdot|l_j, \kappa_V)$ is the univariate von Mises probability density function with mean l_j and precision κ_V. To simulate participants from Group A-V, the model is identical except that visual sequences $V = \{\vec{v}_1, \ldots, \vec{v}_N\}$ are replaced with auditiory sequences $A = \{\vec{a}_1, \ldots, \vec{a}_N\}$, and visual precision κ_V in Eq. 4 is replaced with auditory precision κ_A.

Ideally, we would insert the prior distribution and likelihood function into Bayes' rule (1) to compute the posterior distribution over multisensory representations. Unfortunately, computing the posterior distribution in this manner is intractable, and thus we performed numerical simulations to search the space of multisensory representations. Specifically, we used a tree-based Monte Carlo Markov chain (MCMC) algorithm—a type of Metropolis-Hastings algorithm—based on the algorithm in Goodman et al. (2008).

The algorithm was initialized with a random multisensory representation by drawing a random derivation from the PCFG. This random representation was used as the current hypothesized program, also known as the current state of the Markov chain, at iteration 1. At each subsequent iteration, a proposal program was generated and compared to the current hypothesized program. Proposals were generated as follows. A derivation of a program can be represented by a tree in which internal nodes represent non-terminals and leaf nodes represent terminals. A proposal program was formed by randomly perturbing the current program. A node from the derivation tree of the current program was randomly selected. The subtree below this node was deleted. Non-terminals in the remaining tree were then expanded using random choices of productions from the PCFG.

Finally, a choice was made between the proposal and current program based on the Metropolis-Hastings acceptance function. The proposal was accepted, and thus became the new current program, with probability equal to the minimum of 1 and:

$$\frac{p(V|R')\ |R'|}{p(V|R)\ |R|} \tag{5}$$

where R' is the proposal, $|R'|$ is the number of non-terminals in the derivation of R', R is the current program, and $|R|$ is the number of non-terminals in the derivation of R.[7] This process of randomly generating a proposal and stochastically choosing between the proposal and the current program was repeated for many iterations. The current programs from the final iterations (presumably following convergence of the algorithm) are samples from the posterior distribution over multisensory representations.

Simulation results

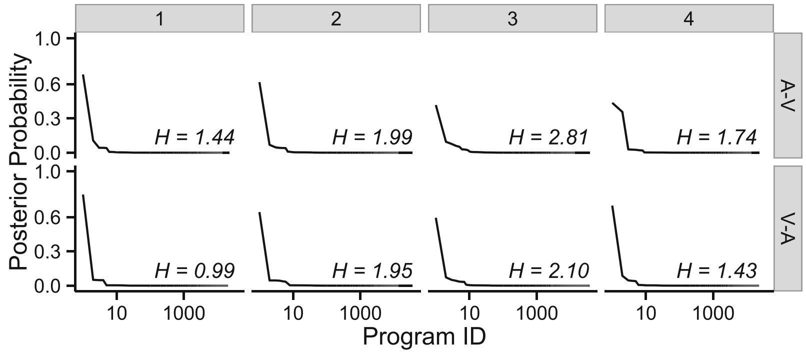
We trained the model in a manner analogous to the way that participants in our experiment were trained. Recall that the experiment had 2 groups (Groups A-V and V-A), with 9 participants per group. Each participant was trained with exemplars from 4 categories. Correspondingly, our model included 2 groups of simulations, one group for auditory training and the other for visual training, with 9 participant-level simulations per group. Each participant-level simulation consisted of 4 category-level simulations. A category-level simulation used the same sensory modality as its corresponding participant, and the same number of exemplars from a category as was observed by this participant during the experiment. For example, consider a category-level simulation corresponding to Category 1, Participant 1, Group A-V. This simulation was conducted using the same number of exemplars from Category 1 as were heard by Participant 1 in Group A-V during the experiment. To mimic sensory noise in our simulations, each location in an exemplar (i.e., a sequence of locations) was perturbed by adding a random number drawn from a von Mises distribution to the location. This was accomplished using the vision-specific or auditory-specific forward models described above.

Each category-level simulation was run for 150,000 iterations of the MCMC algorithm. Samples from the first 100,000 iterations were excluded as burn-in. Samples from the remaining 50,000 iterations were thinned to a set of 5,000 samples to reduce autocorrelations between samples. This set of 5,000 samples is referred to as the category-level simulation's posterior sample.

Posterior distributions

Figure 5 illustrates our results. The first and second rows correspond to Groups A-V and V-A, respectively. The four columns correspond to categories 1-4. Each graph shows the posterior probabilities based on the category-level simulations for a given group and category (there are 9 participant-level simulations per group, and thus there are 9 category-level simulations for a given group and category). The horizontal axis of a graph gives a program identification number. Each program appearing in the posterior sample was assigned a unique ID based on the rank of its posterior probability (the program with the largest probability was numbered 1, the program with the next largest probability was numbered 2, and so on). The vertical axis gives the posterior probability of a program in the posterior sample. For a given group and category, the probability distribution over programs was calculated as follows. For each category-level simulation, we first calculated each program's unnormalized posterior score—defined as the product of a program's likelihood score and

[7] As shown by Goodman et al. (2008), including the number of non-terminals terms, $|R|$ and $|R'|$, in the acceptance function ensures the detailed balance condition of the MCMC algorithm.



MAP

```
L1: init(k)
L2: next(k)
L3: go to L2
```

```
L1: init(k)
L2: next(k)
L3: next(k)
L4: go to L1
```

```
L1: init(k)
L2: prev(k)
L3: prev(k)
L4: prev(k)
L5: go to L1
```

```
L1: init(k)
L2: next(next(k))
L3: prev(k)
L4: go to L2
```

Next-to-MAP (A-V)

```
L1: init(k)
L2: k
L3: next(next(k)
L4: go to L2
```

```
L1: init(k)
L2: k
L3: next(k)
L4: go to L1
```

```
L1: init(k)
L2: prev(k)
L3: prev(prev(k))
L4: k
L5: go to L1
```

```
L1: init(k)
L2: next(k)
L3: k
L4: go to L2
```

Next-to-MAP (V-A)

```
L1: init(prev(k))
L2: next(k)
L3: go to L2
```

```
L1: init(k)
L2: next(init(k))
L3: next(k)
L4: prev(init(k))
L5: go to L2
```

```
L1: init(k)
L2: prev(k)
L3: prev(k)
L4: prev(prev(k))
L5: go to L1
```

```
L1: init(k)
L2: next(k)
L3: k
L4: go to L2
```

Fig. 5 First and second rows show posterior probabilities based on the category-level simulations (see text for details). The horizontal axis of each graph gives a program identification number, and the vertical axis gives the posterior probability of a program in the posterior sample. The entropy (*H*) of each posterior distribution is shown in each graph. The third row shows the program with the highest posterior probability (i.e., the MAP estimate) for each category when samples are combined across all participant-level simulations. The model shows modality invariance as evidenced by the fact that MAP estimates are identical for simulations with auditory and visual training. The fourth and fifth rows show the program with the second-highest posterior probability on the basis of auditory and visual training (Groups A-V and V-A, respectively)

its prior—and then normalized these scores. The normalized scores are a posterior probability distribution over programs for a given category-level simulation. These scores were then averaged across category-level simulations to arrive at the final estimate of a program's posterior probability. The entropy (denoted *H* and measured in bits), an information-theoretic measure of uncertainty (Cover & Thomas, 1991), of each posterior distribution is shown in each graph.

There are several important features of these data. First, posterior distributions are peaked around a single program. Although the hypothesis space of possible programs is infinite (i.e., the probabilistic context-free grammar can generate an infinite number of programs), our results indicate that only a small number of these programs have significant posterior probability for each group and category.

Second, the model shows perfect modality invariance. The third row of Fig. 5 shows the program with the greatest posterior probability for each category when samples are combined across all participant-level simulations. These programs are the model's maximum a posteriori (MAP) estimates of the multisensory representations. Critically, the MAP estimates for each category are identical for simulations of visual and auditory training groups. That is, the model learns the same program regardless of the sensory modality used to perceive a category's training exemplars. Moreover, for all categories, the MAP estimate is correct, meaning that it is identical to the actual program used to generate the exemplars. This result indicates that our model is very effective at learning multisensory representations from sensory data.

The fourth and fifth rows of Fig. 5 show programs with the second-highest posterior probabilities (referred to as 'Next-to-MAP' in the figure) with auditory and visual training, respectively. In some cases, the second-highest scoring program is correct but not the most parsimonious solution (e.g., compare the category 1 MAP estimate and the second-highest scoring Group V-A program). More often, a second-highest scoring program produces sequential patterns that are similar, but not identical, to exemplars from the program's corresponding category.

Lastly, the entropies (i.e., uncertainties) of the posterior distributions follow an interesting pattern. As expected, entropies are higher for auditory training than for visual training. Interestingly, entropies are lowest for Category 1, highest for Category 3, and have intermediate values for Categories 2 and 4. This result suggests that learning about Category 1 should be easiest, learning about Category 3 should be hardest, and learning about Categories 2 and 4 should have intermediate levels of difficulty. This result is consistent with our experimental data (see the right graph in Fig. 3).

Training and test categorization performances

We computed the model's categorization performances as follows. Consider the model's performance during visual training. For the moment, we focus on one participant-level simulation consisting of 4 category-level simulations. For each of these category-level simulations, we calculated the MAP estimate of a multisensory representation (i.e., for each category, we found the program with the highest frequency in the posterior sample). These MAP representations may be regarded as (point estimates of) the category representations acquired by a participant-level simulation. We used them to classify individual training exemplars. Given an exemplar, we computed a posterior score for each MAP representation based solely on that exemplar. The computation of this score used the prior distribution and likelihood function described above (Section "Learning and cross-modal transfer"). The categorization response was taken to be the category corresponding to the MAP representation with highest posterior score. This process was repeated for each of the 36 visual exemplars (9 exemplars × 4 categories) used in a participant-level simulation. Analogous computations were performed during visual testing and during auditory training and testing (as in the experiment, testing included 14 exemplars × 4 categories).

Figure 6 shows the model's categorization performances. The left panel illustrates the training and test results. On average, the participant-level simulations corresponding to Group V-A performed at more than 95 % correct on visual training exemplars. When tested with auditory test exemplars, the simulations showed excellent cross-modal transfer, performing at nearly 85 % correct (recall that chance performance is 25 %). Participant-level simulations corresponding to Group A-V performed at 85 % correct on auditory training items, and more than 90 % correct on visual test items, meaning that this group too showed excellent cross-modal transfer. We emphasize the match between our experimental (left panel of Fig. 3) and modeling (left panel of Fig. 6) results.

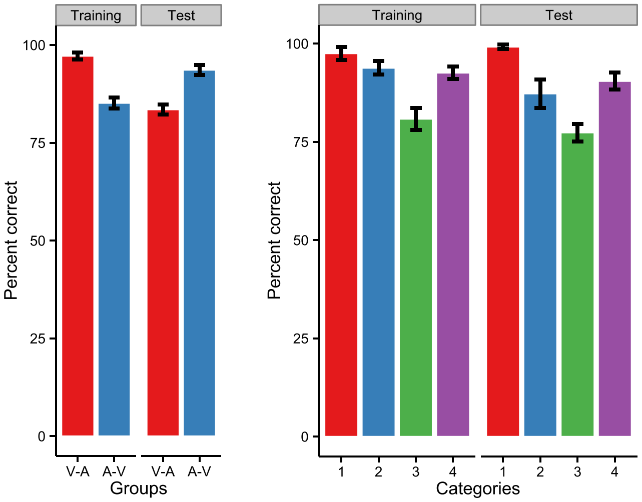


Fig. 6 Modeling results presented in the same format as our experimental results (Fig. 3). (*Left*) Average performances of the participant-level simulations on the training and test exemplars for simulations corresponding to Groups V-A and A-V (error bars indicate standard errors of the means). (*Right*) Participant-level simulations' training and test performances when trials are sorted by the category of the sequence observed on the trial

The right panel of Fig. 6 shows the participant-level simulations' training and test performances when trials are sorted by sequence category. This type of analysis was discussed above in the context of our experimental data (right panel of Fig. 3), and is useful because it allows us to examine the relative ease of correctly classifying exemplars from each category. Recall that experimental participants performed best with exemplars from Category 1, worst with exemplars from Category 3, and at intermediate levels with exemplars from Categories 2 and 4. Does our model show this same rank ordering of category difficulty? The answer is yes. Based on the participant-level simulations' average performances, the rank ordering of the category difficulties parallel the behavioral results.[8] A Friedman test revealed a statistically significant rank ordering of the categories based on training ($p < 0.01$) and test ($p < 10^{-4}$) block responses.

Role of the prior in accounting for participants' performances

When considering our model, we would like to evaluate the role of the prior probability distribution. To do so, we could attempt to develop an alternative model that does not

[8] Rather than use MAP estimates of multisensory representations, an alternative strategy is to randomly sample from the posterior distribution over representations. This alternative strategy provides very similar performances to those shown in Fig. 6. This result was expected because posterior distributions tend to have small variances.

include a prior (i.e, a model in which all multisensory representations are equally probable). In our case, however, this would be difficult to do, particularly when training the model to acquire multisensory representations. As described above, our model learns multisensory representations via an MCMC algorithm in which proposals are generated at each iteration by sampling from the prior distribution. This is where a significant problem arises. If we did not have a prior distribution, then where would proposals come from? Recall that sampling from the prior occurs through the use of the production rules in the PCFG. Even if we could devise a scheme in which all derivations from the PCFG are equally probable (it is not clear that we could), then this would likely lead to other challenges, such as creating MCMC algorithms that converge.

The model uses the prior distribution both when acquiring multisensory representations and when using these representations to categorize an exemplar. Given that it may be impossible to ignore the prior during the acquisition stage, we decided to focus on what would happen if we ignored the prior during the categorization stage. When categorizing an exemplar, our model calculated a posterior score for each multisensory MAP representation, and labeled the exemplar based on the representation with the highest score. We wondered what would happen if we did not calculate a posterior score but, rather, calculated a likelihood score. That is, what if MAP representations were evaluated based solely on the likelihood function, ignoring the prior distribution?

Figure 7 shows the results. Clearly, the model's results are now less similar to the experimental results. For example, differences between training and test performances are now greatly reduced. Furthermore, the rank ordering of category difficulties observed in the behavioral data and in the original model's data is absent from this model's data. Perhaps the most striking aspect of this model's data is that the performances are so high. That is, ignoring the prior distribution when evaluating multisensory MAP representations leads to better performances. At the same time, it also makes the model's performances less similar to experimental participants' performances, suggesting that people, like our model, might also have a bias toward “simpler programs”. Future work will need to address this hypothesis.

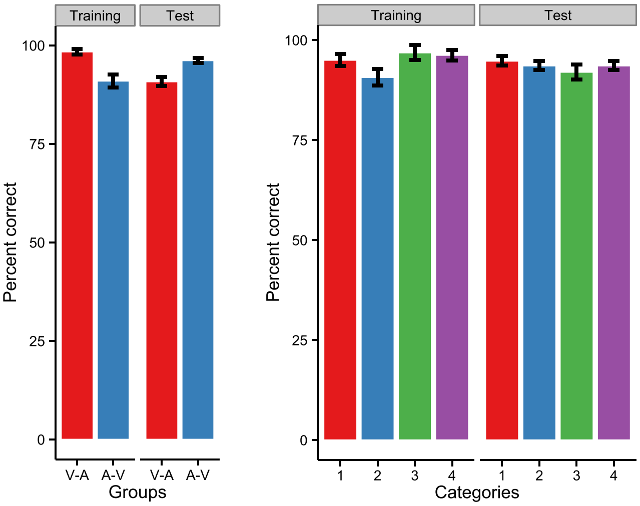


Fig. 7 Results from the model in which multisensory MAP representations are evaluated based on a likelihood score, not a posterior score

Discussion

In summary, our goal has been to use the Multisensory Hypothesis to better understand how people acquire and use multisensory representations to facilitate transfer of knowledge across sensory modalities. We conducted an experiment evaluating whether people transfer sequence category knowledge across auditory and visual domains. Our experimental data clearly indicate that we do. We then developed a computational model accounting for our experimental results. To our knowledge, this is among the first formulations of the Multisensory Hypothesis that has been explicitly defined and implemented (also see Yildirim & Jacobs, 2013). Because our model demonstrates how the acquisition and use of amodal, multisensory representations can underlie cross-modal transfer of knowledge, and because our model accounts for subjects' performances, our work lends credence to the Multisensory Hypothesis. Overall, our work suggests that people automatically extract and represent objects' and events' intrinsic properties, and use these properties to process and understand the same (and similar) objects and events when they are perceived through novel sensory modalities.

pLOT approach beyond higher-level cognition Multisensory representations lie at the core of our computational model. An unusual aspect of the model is that these representations are characterized as computer programs, and programs are learned via Bayesian inference. As discussed above, our work contributes to the emerging pLOT perspective. Symbolic and statistical approaches to cognitive modeling often have complementary strengths and weaknesses. A strength of symbolic approaches is their representational expressiveness which comes from their use of highly structured, compositional data structures. However, symbolic approaches are often “brittle” (i.e., they often fail in uncertain environments) and often have limited learning capabilities. In contrast, statistical approaches tend to be robust in the sense that they often work well despite uncertainty. In addition, these approaches can excel at inference and learning, especially when using new computational techniques developed in the past 25 years (e.g., new Monte Carlo sampling methods or variational approximations). However, statistical approaches often require highly structured prior distributions or likelihood functions to work well (Tenenbaum, Kemp, Griffiths, & Goodman, 2011). By combining the strengths of symbolic and statistical approaches, the pLOT perspective may offer a unifying framework for thinking about many aspects of human cognition. Our work extends the small (but hopefully growing) literature on the pLOT modeling approach. To our knowledge, our model is among the first pLOT models to address the domain of human perception.

Extensions of the current model Our model is an instance of a "single cause" model because it assumes that visual and auditory signals arise from a single source (i.e, a sequence of events). In the real world, however, visual and auditory signals sometimes arise from the same source and other times arise from different sources. People are able to learn if different sensory signals should be attributed to the same or different underlying causes. Future extensions of the current model will need to learn this too (Körding et al., 2007).

Future extensions will also need to consider how the model can be scaled to larger, more realistic scenarios. In more realistic settings, richer sets of representational primitives will be needed, as well as more sophisticated forward models. We are encouraged by the fact that researchers are developing advanced software for perceptual (e.g., visual, auditory) rendering, for simulating the kinematics and dynamics of robots, and for simulating dynamic interactions among objects. Cognitive scientists can build larger, more realistic models by using these software packages as forward models in their models of human perception, motor control, and intuitive physics (see Battaglia, Hamrick, & Tenenbaum, 2013; Yildirim & Jacobs, 2013).

Computational and representation/algorithm levels of analysis Cognitive models are often classified based on whether they contribute to computational or representational/algorithmic levels of analysis (Marr, 1982). We believe that our model currently makes a contribution at the computational level and may, in the future, make a contribution at the representational/algorithmic level. At the computational level, our model defines optimal performance on our experimental task (given the assumptions of the model; see Jacobs & Kruschke, 2011). Therefore, it can be used as a benchmark to evaluate subjects' performances. Subjects performed correctly on about 70-75 % of final training and test trials. In addition, they performed best on exemplars from Category 1, worst on exemplars from Category 3, and at intermediate levels on exemplars from Categories 2 and 4. Are these performances good or bad? By comparing subjects' performances with those of the computational model, we see that subjects' performances are similar to those of the model, though subjects are moderately less proficient. This indicates that subjects performed well, but that there was still room for improvements in these performances. The gap between subjects' performances and the model's performances may have been due to our training procedures. Future work will need to investigate this issue.

Our model can potentially be used as a starting point for a new model intended to faithfully capture people's psychological operations and representations underlying cross-modal transfer. In particular, our simulation results make a case in favor of the use of compositional representations for understanding multisensory perception. Future research will need to study the psychological plausibility and the detailed role of compositional representations in human perception.

Acknowledgments We thank S. Piantadosi for many helpful discussions, and the editor and anonymous reviewers for constructive comments on this manuscript. This work was supported by research grants from the National Science Foundation (DRL-0817250, BCS-1400784) and the Air Force Office of Scientific Research (FA9550-12-1-0303).

References

Alais, D., & Burr, D. (2004). The ventriloquist effect results from near-optimal bimodal integration. *Current Biology*, *14*, 257–262.

Barsalou, L. W. (2008). Grounded cognition. *Annual Review of Psychology*, *59*, 617–645.

Battaglia, P. W., Hamrick, J. B., Tenenbaum, J. B. (2013). Simulation as an engine of physical scene understanding. *Proceedings of the National Academy of Sciences USA*, *110*, 18327–18332.

Battaglia, P. W., Jacobs, R. A., Aslin, R. N. (2003). Bayesian integration of visual and auditory signals for spatial localization. *Journal of the Optical Society of America A*, *20*, 1391–1397.

Bo, J., & Seidler, . D. (2010). Spatial and symbolic implicit sequence learning in young and older adults. *Experimental Brain Research*, *201*, 837–851.

Calvert, G. A., Bullmore, E. T., Brammer, M. J., Campbell, R., Williams, S. C. R., McGuire, P. K., ... David, A. S. (1997). Activation of auditory cortex during silent lipreading. *Science*, *276*, 593–596.

Calvert, G. A., Spence, C., Stein, B. E. (2004). *The Handmisc of Multisensory Processes*. Cambridge, MA: MIT Press.

Cleeremans, A., & McClelland, J. L. (1991). Learning the structure of event sequences. *Journal of Experimental Psychology: General*, *120*, 235–253.

Cover, T. M., & Thomas, J. A. (1991). *Elements of Information Theory*. Wiley: New York.

de Gelder, B., & Vroomen, J. (2000). The perception of emotions by ear and by eye. *Cognition and Emotion*, *14*, 289–311.

Deroost, N., & Soetens, E. (2006). Spatial processing and perceptual sequence learning in SRT tasks. *Experimental Psychology*, *53*, 16–30.

Elman, J. L. (1990). Finding structure in time. *Cognitive Science*, *14*, 179–211.

Fiser, J., & Aslin, R. (2002). Statistical learning of higher-order temporal structure from visual shape sequences. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *28*, 458–467.

Fodor, J. A. (1975). *The Language of Thought*. Cambridge, MA: Harvard University Press.

Goodman, N. D., Tenenbaum, J. B., Feldman, J., Griffiths, T. L. (2008). A rational analysis of rule based concept learning. *Cognitive Science*, *32*, 108–154.

Gureckis, T. M., & Love, B. C. (2010). Direct associations or internal transformations? Exploring the mechanisms underlying sequential learning behavior. *Cognitive Science*, *34*, 10–50.

Hunt, R., & Aslin, R. (2001). StatisticalStatistical learning in a serial reaction time task: Access to separable statistical cues by individual learners. *Journal of Experimental Psychology: General*, *130*, 658–680.

Hunt, R. H., & Aslin, R. N. (2010). Category induction via distributional analysis: Evidence from a serial reaction time task. *Journal of Memory and Language*, *62*, 98–112.

Ito, M. (2008). Control of mental activities by internal models in the cerebellum. *Nature Reviews Neuroscience*, *9*, 304–313.

Jacobs, R. A., & Kruschke, J. K. (2011). Bayesian learning theory applied to human cognition. *Wiley Interdisciplinary Reviews: Cognitive Science*, *2*, 8–21.

Jordan, M. (1986). Attractor dynamics and parallelism in a connectionist sequential machine. In *Proceedings of the Twelfth Annual Conference of the Cognitive Science Society* (pp. 531–546). Hillsdale, NJ: Erlbaum.

Körding, K. P., Beierholm, U., Ma, W. J., Quartz, S., Tenenbaum, J., Shams, L. (2007). Causal inference in multisensory perception. *PLoS ONE*, *2*(9), e943.

Lehmann, S., & Murray, M. M. (2005). The role of multisensory memories in unisensory object discrimination. *Cognitive Brain Research*, *24*, 326–334.

Liang, M., Mouraux, A., Hu, L., Iannetti, G. D. (2013). Primary sensory cortices contain distinguishable spatial patterns of activity for each sense. *Nature Communications*, *4*, 1979.

Marr, D. (1982). Vision: A Computational Investigation Into the Human Representation and Processing of Visual Information.

McCallum, A. R. (1996). Learning to use selective attention and short-term memory in sequential tasks. *From Animals to Animats: Proceedings of the Fourth International Conference on Simulation of Adaptive Behavior*.

McClelland, J. L., & Patterson, K. (2002a). 'Words or Rules' cannot exploit the regularity in exceptions. *Trends in Cognitive Sciences*, *6*, 464–465.

McClelland, J. L., & Patterson, K. (2002b). Rules or connections in past-tense inflections: What does the evidence rule out?. *Trends in Cognitive Sciences*, *6*, 465–472.

Miall, R. C., & Wolpert, D. M. (1996). Forward models for physiological motor control. *Neural Networks*, *9*, 1265–1279.

Pascual-Leone, A., & Hamilton, R. (2001). The metamodal organization of the brain. *Progress in Brain Research*, *134*, 427–445.

Pekkola, J., Ojanen, V., Autti, T., Jääskeläinen, I. P., Möttönen, R., Tarkiainen, A., Sams, M. (2005). Primary auditory cortex activation by visual speech: An fMRI study at 3T. *NeuroReport*, *16*, 125–128.

Piantadosi, S. T., Tenenbaum, J. B., Goodman, N. D. (2012). Bootstrapping in a language of thought: A formal model of numerical concept learning. *Cognition*, *123*, 199–217.

Pinker, S., & Ullman, M. T. (2002a). The past and future of the past tense. *Trends in Cognitive Sciences*, *6*, 456–463.

Pinker, S., & Ullman, M. T. (2002b). Combination and structure, not gradedness, is the issue. *Trends in Cognitive Sciences*, *6*, 472–474.

Quiroga, R. Q. (2012). Concept cells: The building blocks of declarative memory functions. *Nature Reviews Neuroscience*, *13*, 587–597.

Quiroga, R. Q., Kraskov, A., Koch, C., Fried, I. (2009). Explicit encoding of multimodal percepts by single neurons in the human brain. *Current Biology*, *19*, 1308–1313.

Simon, H. A., & Kotovsky, K. (1963). Human acquisition of concepts for sequential patterns. *Psychological Review*, *70*, 534–546.

Stein, B. E. (2012). *The New Handmisc of Multisensory Processing*. Cambridge, MA: MIT Press.

Tanabe, H. C., Honda, M., Sadato, N. (2005). Functionally segregated neural substrates for arbitrary audiovisual paired-association learning. *Journal of Neuroscience*, *25*, 6409–6418.

Tenenbaum, J. B., Kemp, C., Griffiths, T. L., Goodman, N. D. (2011). How to grow a mind: Statistics, structure, and abstraction. *Science*, *331*, 1279–1285.

Tian, X., & Poeppel, D. (2010). Mental imagery of speech and movement implicates the dynamics of internal forward models. *Frontiers in Psychology*, *1*, 1–23.

Ullman, T. D., Goodman, N. D., Tenenbaum, J. B. (2012). Theory learning as stochastic search in the language of thought. *Cognitive Development*, *27*, 455–480.

von Kriegstein, K., & Giraud, A.-L. (2006). Implicit multisensory associations influence voice recognition. *PLoS Biology*, *4*, e326.

Wallraven, C., Bülthoff, H. H., Waterkamp, S., van Dam, L., Gaißert, N. (2014). The eyes grasp, the hands see: Metric category knowledge transfers between vision and touch. *Psychonomic Bulletin &amp; Review*, in press.

Yildirim, I., & Jacobs, R. A. (2013). Transfer of object category knowledge across visual and haptic modalities: Experimental and computational studies. *Cognition*, *126*, 135–148.