8. H.R. Markus and S. Kitayama, Culture and the self: Implications for cognition, emotion, and motivation, *Psychological Review, 98,* 224–253 (1991); H.C. Triandis, The self and social behavior in differing cultural contexts, *Psychological Review,* 96, 506–520 (1989); R.A. Shweder and M.A. Sullivan, Cultural psychology: Who needs it? *Annual Review of Psychology, 44,* 497–523 (1993).

9. C.D. Ryff, Y.H. Lee, and K.C. Na, *Through the lens of culture: Psychological well-being at midlife,* unpublished manuscript, University of Wisconsin, Madison (1995).

10. R.M. Hauser, W.H. Sewell, J.A. Logan, T.S. Hauser, C.D. Ryff, A. Caspi, and M.M. MacDonald, The Wisconsin Longitudinal Study: Adults as parents and children at age 50, *IASSIST Quarterly, 16,* 23–38 (1992); S.M. Heidrich and C.D. Ryff, The role of social comparisons processes in the psychological adaptation of elderly adults, *Journal of Gerontology, 48,* P127–P136 (1993); Ryff and Essex, note 1; C.D. Ryff and M.J. Essex, The interpretation of life experience and well-being: The sample case of relocation, *Psychology and Aging, 7,* 507–517 (1992); C.D. Ryff, Y.H. Lee, M.J. Essex, and P.S. Schmutte, My children and me: Mid-life evaluations of grown children and of self, *Psychology and Aging, 9,* 195–205 (1994); S. Tweed and C.D. Ryff, Adult children of alcoholics: Profiles of wellness and distress, *Journal of Studies on Alcohol, 52,* 133–141 (1991).

Perceiving Talking Faces

Dominic W. Massaro and Michael M. Cohen

No one doubts the importance of the face in social interactions, but people seldom think of it as playing much of a role in verbal communication. A number of observations suggest otherwise, though: Many people dislike talking over the telephone and are irritated by poorly dubbed foreign films. Some people even comment that they hear the television better with their glasses on. Children born blind learn some speech distinctions more slowly than their sighted cohorts. It has been well known for some time that the deaf and hearing impaired can make valuable use of lipreading, which is better termed speechreading, but more recently investigators have shown that even people with normal hearing are greatly influenced by the visible speech in face-to-face communication. Our research is aimed at understanding how people perceive speech by both ear and eye.

Dominic W. Massaro is Professor of Psychology and **Michael M. Cohen** is Research Associate in the Department of Psychology, University of California, Santa Cruz. Address correspondence to Dominic W. Massaro, Department of Psychology, University of California, Santa Cruz, CA 95064; e-mail: massaro@fuzzy.ucsc.edu; WWW URL: http://mambo.ucsc.edu/psl/dwm.html.

PERCEIVING SPOKEN LANGUAGE

Although people take understanding speech for granted, it is an amazing accomplishment. No computer has been programmed to understand speech as well as a 3-year-old child. One reason people are such experts is their ability to use many different cues to disambiguate a message. Some stimulus cues are contained in the speech signal, and others are present in the situational and linguistic context. An example of an auditory cue is the /s/ in *sin;* this sound has a particular noise quality that differs from that of the /š/ in *shin.* Contextual cues from the word and sentence can also be important. For example, if the /s/ segment in *legislature* is replaced by a musical tone, a listener may still perceive the word as intact. Even less of the word is necessary for recognition when it is spoken in a sentence, such as "The governor gave an address to the state _______."

In face-to-face communication, there are also important cues available from the face, lips, and tongue of the speaker. Of course, hearing-impaired persons benefit greatly from visible speech, but even individuals with normal hearing are influenced by these visible cues. If you make an auditory tape of the nonsense sentence "My bab pop me poo brive," and dub it onto a videotape of someone saying, "My gag kok me koo grive," a viewer will be likely to hear, "My dad taught me to drive." In this example, first created by Harry McGurk,[1] the nonsense from each of the two modalities was selected to approximate the meaningful sentence. Auditory "brive" provides strong support for "brive" but also some support for "drive." Similarly, visual "grive" provides support for both "grive" and "drive," and very little support for "brive." In this case, "drive" is the best interpretation because it has substantial support from both the auditory and the visual sources of information. A similar analysis can be given for the other segments that have conflicting auditory and visual information. The perceiver naturally combines the auditory and visual sentences into something meaningful because the auditory and visual inputs are both reasonably consistent with the meaningful sentence.

Although this example involves the interpretation of a sentence, the research we present addresses more directly the perception of a single speech segment without meaning. Our research is carried out in the framework of a fuzzy logical model of perception (FLMP).[2] The central assumption of this approach is that perceiving speech is fundamentally a pattern recognition problem.

Published by Cambridge University Press

Within this framework, perceptual recognition of speech patterns, as of other objects and events, depends on three processes: feature evaluation, feature integration, and decision (see Fig. 1). The temporal occurrence of these processes is necessarily successive, although they overlap in time. Spoken language is transduced by the sensory systems, which make available a set of sensory primitives, called sensory cues or features. As members of a linguistic community, people have knowledge about what segments of speech occur in their language. Each segment of language is represented in memory by a prototype defined in terms of its ideal features.

In the FLMP, auditory and visual features of spoken language are evaluated to determine the degree to which each feature supports each prototype. In contrast to most models of speech perception, the features are assumed to provide continuous, rather than discrete, information. In other words, a particular feature may be said to support a particular prototype to some degree, rather than simply to support or not support that prototype. Given that both the auditory and the visual features are evaluated in terms of support for alternative prototypes, the integration process can easily combine this information to give an overall degree of support for each prototype. Because both auditory and visual features are processed, the model provides a natural account of the joint influence of auditory and visual information in speech perception.

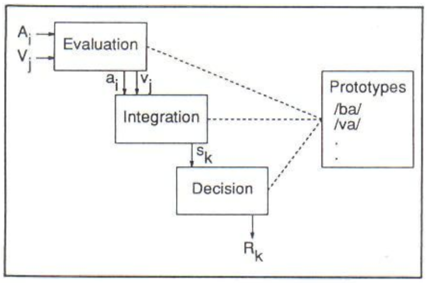

Fig. 1. Schematic representation of the three processes involved in perceptual recognition. The three processes are shown to proceed left to right in time to illustrate that they are necessarily successive but overlapping. These processes make use of prototypes stored in long-term memory. The sources of information are represented by uppercase letters: Auditory information is represented by A_i and visual information by V_j. The evaluation process transforms these sources of information into psychological values (indicated by lowercase letters a_i and v_j). These sources are then integrated to give an overall degree of support, s_k, for a given speech alternative k. The decision operation maps the outputs of integration into some response alternative, R_k. The response can take the form of a discrete decision or a rating of the degree to which the alternative is likely.

The final step in the model is the decision process. This process makes some judgment on the basis of the relative support for the relevant prototypes.

EXPANDED FACTORIAL DESIGN

Given this theoretical framework, our study of speech perception usually involves the independent variation of several sources of information. A particularly valuable experimental paradigm is to vary the two modalities independently in what is called an expanded factorial design. Figure 2 illustrates the design used in one study in which each of four auditory syllables was combined with each of four visible syllables (bimodal presentations). In addition, each of the syllables was presented only auditorily and only visually (unimodal presentations). The goal of this type of study is to determine how the separate sources of information are processed together to achieve speech perception. The expanded factorial design provides a strong test of quantitative models because it has both unimodal and bimodal conditions. Each candidate model must describe the relationship between unimodal and bimodal performance.

A TALKING HEAD

To create synthetic visible speech, we modified and extended a system first developed by Fredric Parke.[3] A fairly realistic animated face is composed of many polygons joined together and controlled by a set of parameters. The face is modeled as a polyhedral object composed of about 900 small triangles arranged in three dimensions and joined together at the edges. The left panel of Figure 3 shows a framework rendering of this model. To achieve a natural appearance, the surface is smooth shaded (shown in the right panel of Fig. 3). The face is animated by altering the location of various points in the grid under the control of about 65 parameters. The face has eyes, nose, mouth, teeth, and a tongue. Each speech segment is defined in a table according to target values for 18 control parameters and segment duration. These control parameters include rotating and thrusting the jaw, varying the horizontal width of the mouth, protruding the lips, moving the corners of the mouth, tucking the lower lip under the upper teeth, raising the upper and lower lips, and varying the angle, width, length, and thickness of the tongue. The nonspeech parameters control the eyes, eyebrows, and position, size, and other aspects of the face. The animation is carried out in real time on a Silicon Graphics Inc. Crimson-VGX computer.

We used synthetic visible speech and natural audible speech to generate the consonant-plus-vowel (CV) syllables /ba/, /va/, /ða/, and /da/. Figure 4 shows a view of the talking head at the onset of articulation for each syllable. Using an expanded factorial design, the four syllables were presented auditorily, visually, and bimodally. For the bimodal presentation, each audible syllable was presented with each visible syllable for a total of 16 (4 × 4) unique conditions. Twelve of the bimodal syl-

		Visual				
		/ba/	/va/	/ða/	/da/	None
	/ba/					
	/va/					
Auditory	/ða/					
	/da/					
	None					

Fig. 2. Expanded factorial design with four auditory syllables crossed with four visual syllables. Note that the consonant /ð/ is pronounced like the "th" in "the."

lables had inconsistent auditory and visual information. These conditions are necessary to achieve an informative picture of how these two speech modalities are processed. More generally, the goal of our research is to determine a theoretical description that can describe or explain performance on the bimodal conditions as a function of performance on the unimodal conditions. Most important, this experimental design allows us to control and manipulate the audible and visible speech independently of one another.

The 10 participants in the experiment were instructed to watch the talking head and listen on each trial and to indicate the consonant that was spoken. The subjects made their responses by pressing labeled keys on a computer keyboard: "b," "v," "th," or "d" alone for a single consonant or two keys successively for a consonant cluster (e.g., if the subject heard /bda/). All of the 24 test syllables were randomized and presented 20 times each for identification. The mean observed proportion of times each response was given was computed for each subject for each of the 24 test syllables by pooling across all 20 experimental trials for each condition.

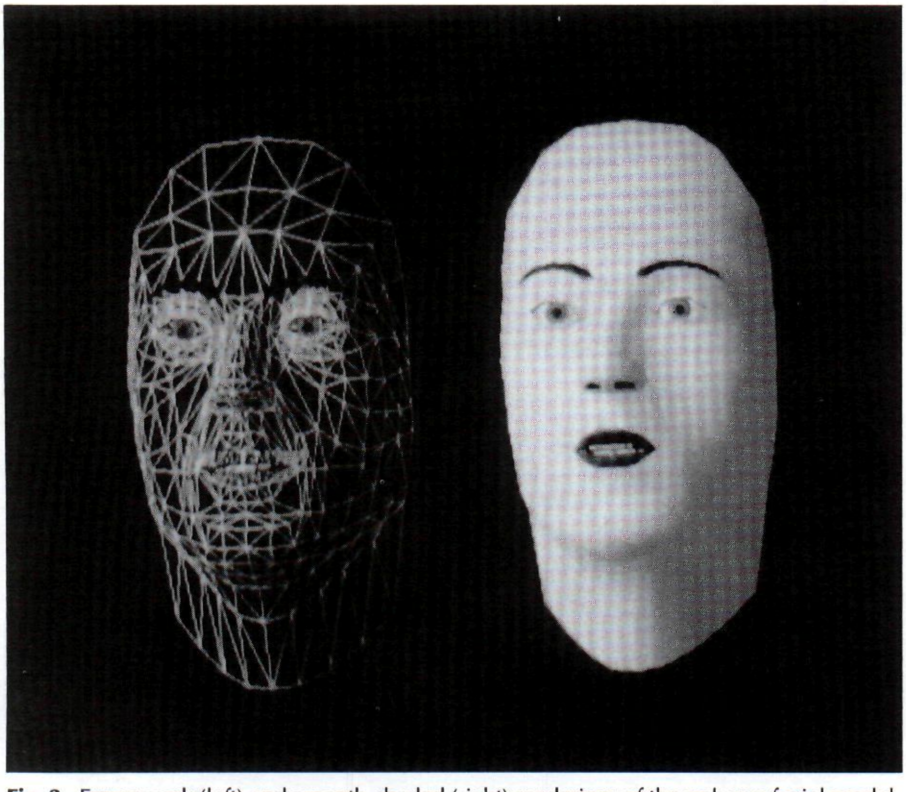

Fig. 3. Framework (left) and smooth-shaded (right) renderings of the polygon facial model.

SPEECH BY EYE AND EAR

In our task, the participants' goal was to perceive what was spoken. We can assess the influence of the auditory and visual speech by evaluating subjects' accuracy with respect to each of these modalities. The left panel in Figure 5 shows average performance scored in terms of accuracy with respect to the visible speech. For unimodal trials, the average correct performance was .89, .91, .78, and .70 for the visible syllables /ba/, /va/, /ða/, and /da/, respectively. Thus, perceivers are fairly good at speechreading these syllables, and the more visually distinctive syllables /ba/ and /va/ are somewhat easier than the others.

Performance scored in terms of accuracy with respect to the auditory modality is given in the right panel of Figure 5. Correct identification averaged .77, .72, .88, and .99 for the unimodal auditory syllables /ba/, /va/, /ða/, and /da/, respectively. The auditory syllables /ða/ and /da/ were perceived more accurately than the syllables /ba/ and /da/. The different levels of performance on the auditory and visual syllables tend to replicate a more general complementarity of these

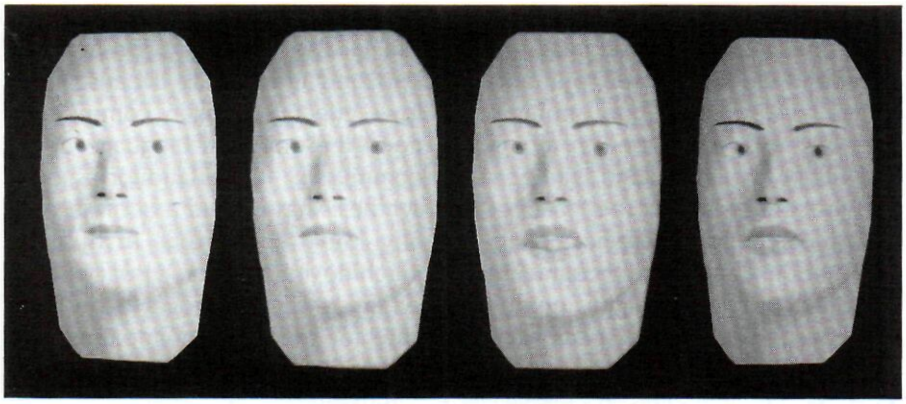

Fig. 4. The facial model at the onset of each of the four syllables tested. From left to right, the lips are closed at the onset of /ba/, much of the lower lip is hidden by the teeth in /va/, the tongue is between the teeth in /ða/ (written as DH), and the mouth is slightly open at the onset of /da/.

two modalities in speech perception. Several syllables easy in the visual modality tend to be difficult in the auditory modality, and vice versa.

Some readers might be surprised that the participants made errors on the unimodal trials. However, people are seldom expected to recognize isolated syllables and usually have the benefit of supplementary contextual cues. In addition, these four syllables are fairly similar to one another and, therefore, easily confused. For example, recently one of us was interviewed over the telephone by *"Tanorama,"* an Italian weekly magazine. The name of the magazine was perplexing because we did not know this magazine or even the meaning of the word. Only later did we learn that the magazine was actually *Panorama*. This confusion between /t/ and /p/ would have been precluded if visible speech had also been available.

Figure 5 also graphs accuracy on bimodal trials for consistent and inconsistent trials; the left panel of Figure 5 is scored with respect to accuracy on the visual syllable, whereas the right panel is scored with respect to the auditory syllable. The results show a large influence of both modalities on performance. Overall performance was more accurate with two sources of consistent information than with just one of these sources of information. Furthermore, given two sources of inconsistent information, performance was poorer than observed in the unimodal conditions.

Figure 5 also reveals that there was a larger influence from the auditory than the visual source of information. Inconsistent auditory information disrupted visual performance more than inconsistent visual information disrupted auditory performance. Similarly, but to a smaller degree, consistent auditory information improved visual performance more than consistent visual information improved auditory performance. This advantage of the auditory over the visual modality is not due to the use of natural auditory and synthetic visual speech. The same result has been observed with natural visual and synthetic auditory speech.[4] More generally, auditory speech is more informative than visual: People can communicate over the telephone but not via silent video.

These results also provide a test of the FLMP. The model was tested against the individual results of each of the 10 participants. The model predicts the actual responses, rather than simply accuracy for one modality or the other. Observed judgments of the bimodal stimuli included 11 different consonants or consonant clusters, counting only those that occurred on more than 0.7% of the bimodal trials. The model was fit to these 11 judgments (plus a 12th "other" category). The predictions of these judgments deviated from the observations by an average of only 0.02. Rather than plot all of the identification judgments and predictions, we simply analyzed their ac-

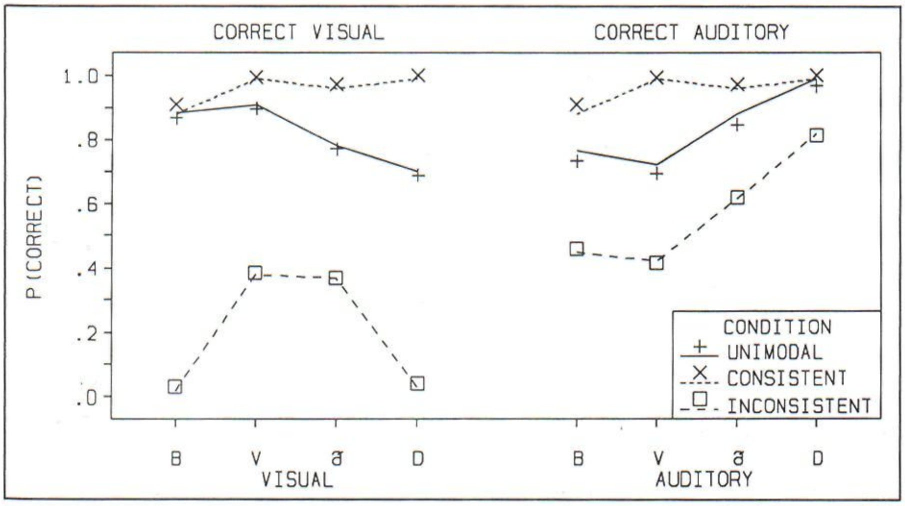

Fig. 5. Average observed and predicted accuracy in identifying syllables. Accuracy with respect to visual information is graphed on the left; accuracy with respect to auditory information is graphed on the right. Proportion correct is graphed for unimodal trials, bimodal trials when the auditory information was consistent with the visual, and bimodal trials when the auditory information was inconsistent with the visual. Points give the observed accuracy, and lines give the predictions of the fuzzy logical model of perception.

curacy to give the average observations and predictions in Figure 5.

It is also worthwhile to mention some interesting response patterns for a few inconsistent auditory-visual combinations. The participants often perceived consonant clusters when the visible speech was articulated more forward in the mouth than the audible speech. Faced with a visual /ba/ and an auditory /va/, subjects responded with "bv" 23% of the time. Visual /ba/ and auditory /da/ produced 21% "bd" judgments. In other cases, the perceptual response differed from both syllables. For example, visual /da/ and auditory /ba/ produced 22% "th" responses, and visual /da/ and auditory /va/ produced "th" responses 51% of the time. The FLMP predicted these judgments accurately.

FLMP VERSUS SPEECH IS SPECIAL

In the space allotted here, we are not able to review and evaluate other extant theories.[5] However, the success of the FLMP weakens a contrasting viewpoint that language and speech involve specialized processes. According to this speech-is-special theory, speech perception cannot be understood in terms of general principles of perception and pattern recognition.[6] One claim is that some processes of speech production must necessarily be engaged in the act of speech perception. Certainly, the information supporting speech perception is specialized in the sense of being unique to speech distinctions in the speaker's language. There is evidence, however, that the processes involved in evaluating and integrating this information are analogous to those in other domains, such as object recognition.

If speech and facial affect are recognized by different specialized processors, then the results in the two domains should be very different from one another. To address this question, we have used our animated face to study how multiple cues in the face are used to recognize affect.[7] This research is important because, as in the case of speech perception, there has been a long-standing belief that recognizing the affect in faces is highly unique and unusual—involving specialized brain areas using holistic and nonanalytic processes.[8] This hypothesis can be adequately tested with our synthetic head, whose parts can be changed independently of one another. The position of the brows and position of the mouth are two important cues for happiness and anger, for example. Using our animated face, we manipulated these two cues independently of one another and asked people to judge the affect of the face. The results indicated that people use both cues to judge affect, and they combine the features in the same manner that they combine speech features. Thus, affect processing is shown to be highly analogous to speech processing: Both are well described by the FLMP. One trademark of both domains and the FLMP is that the influence of a given cue is greatest when the other cues are ambiguous. Although many investigators have argued otherwise, both speech and facial affect appear to be recognized in the same manner as other objects and events are.

APPLICATIONS

One applied value of visible speech is its potential to supplement other (degraded) sources of information. The value of visible speech is not limited to speech perception. It has also been shown that the added dimension of visible speech can facilitate comprehension and memory of spoken language.[9] Visible speech is particularly beneficial in poor listening environments with substantial amounts of background noise. Its use is also important for hearing-impaired individuals because it allows effective spoken communication—the universal language of the community. Just as auditory speech synthesis has proved a boon to visually impaired citizens in human–machine interaction, visual speech synthesis should prove to be valuable for the hearing impaired. For example, synthetic visible speech (simulating a specific person's face by placing a representation of it on the surface of the animated head) can be used in videophones with normal telephone transmission. The addition of visual cues provided by the face would allow individuals previously unable to communicate remotely to do so.

The results of this research can also be used to implement automatic speechreading to enhance speech recognition by machine. If human perceivers achieve robust recognition of speech by using multiple sources of information, the same should be true for machines. Finally, synthetic visible speech has an important part in building synthetic "actors" and should play a valuable role in the exciting new sphere of virtual reality.

Acknowledgments—The research described in this review was supported, in part, by grants from the Public Health Service (PHS R01 DC 00236), the National Science Foundation (BNS 8812728), and the University of California, Santa Cruz. The authors thank Bill Farrar, Lisa Lima, and Anne Shin for reading and commenting on the manuscript.

Notes

1. H. McGurk, Listening with eye and ear (paper discussion), in *The Cognitive Representation of Speech*, T. Myers, J. Laver, and J. Anderson, Eds. (North-Holland, Amsterdam, 1981).

2. D.W. Massaro, *Speech Perception by Ear and Eye: A Paradigm for Psychological Inquiry* (Erlbaum, Hillsdale, NJ, 1987); D.W. Massaro, Multiple book review of *Speech Perception by Ear and Eye: A Paradigm for Psychological Inquiry*, *Behavioral and Brain Sciences, 12*, 741–794 (1989); D.W. Massaro and D. Friedman, Models of integration given multiple sources of information, *Psychological Review, 97*, 225–252 (1990).

[3] D.W. Massaro, Broadening the domain of the fuzzy logical model of perception, in *Cognition: Conceptual and Methodological Issues*, H.L. Pick, Jr., P. Van den Broek, and D.C. Knill, Eds. (American Psychological Association, Washington, DC, 1992).

[4] M.M. Cohen and D.W. Massaro, Synthesis of visible speech, *Behavioral Research Methods, Instrumentation, & Computers, 22,* 260–263 (1990); M.M. Cohen and D.W. Massaro, Modeling coarticulation in synthetic visual speech, in *Models and Techniques in Computer Animation,* N.M. Thalmann and D. Thalmann, Eds. (Springer-Verlag, Tokyo, 1993); M.M. Cohen and D.W. Massaro, Development and experimentation with synthetic visible speech, *Behavioral Research Methods, Instrumentation, & Computers, 26,* 260–265 (1994); F.I. Parke, A model for human faces that allows speech synchronized animation, *Computers and Graphics Journal, 1*(1), 1–4 (1975).

[5] B. de Boysson-Bardies, S. de Schonen, P. Jusczyk, P. McNeilage, and J. Morton, Eds., *Developmental Neurocognition: Speech and Face Processing in the First Year of Life* (Kluwer, Dordrecht, The Netherlands, 1993); C.A. Fowler and D.J. Dekle, Listening with eye and hand: Cross-modal contributions to speech perception, *Journal of Experimental Psychology: Human Perception and Performance, 17,* 816–828 (1991); Q. Summerfield, Visual perception of phonetic gestures, in *Modularity and the Motor Theory of Speech Perception,* I.G. Mattingly and M. Studdert-Kennedy, Eds. (Erlbaum, Hillsdale, NJ, 1991).

[6] A. Liberman and I.G. Mattingly, The motor theory of speech perception revised, *Cognition, 21,* 1–33 (1985); I.G. Mattingly and M. Studdert-Kennedy, Eds., *Modularity and the Motor Theory of Speech Perception* (Erlbaum, Hillsdale, NJ, 1991).

[7] J.W. Ellison and D.W. Massaro, *Evaluating and integrating features in the identification of facial affect,* unpublished manuscript, University of California, Santa Cruz (1994).

[8] S.C. Levine, M.T. Banich, and M.P. Koch-Weser, Face recognition: A general or specific right hemisphere capacity? *Brain and Cognition, 8,* 303–325 (1988).

[9] B. Dodd and R. Campbell, Eds., *Hearing by Eye: The Psychology of Lip-Reading* (Erlbaum, Hillsdale, NJ, 1987).

Early Understanding and Use of Symbols: The Model Model

Judy S. DeLoache

The hallmark of human cognition is symbolization: There is nothing that so clearly distinguishes us from other creatures as our creative and flexible use of symbols. Cultural creations such as writing systems, number systems, maps, and models—to name a few—have enabled human knowledge and reasoning to transcend time and space.

My working definition of an external, artifactual symbol is that it is any entity that someone intends to stand for something other than itself. Note that this definition is agnostic about the nature of symbols; virtually anything can be a symbol, so long as some person intends that it be responded to not as itself, but in terms of what it represents. Adults are so experienced and skilled with symbols and symbolic reasoning that they simply assume that many of the novel entities they encounter will have symbolic import. They appreciate that such entities should be responded to as representations of something other than themselves—and readily do so. My research reveals that children only gradually adopt this assumption. Despite the centrality of symbolization in human cognition and communication, young children are very conservative when it comes to detecting and reasoning about symbol–referent relations.

Judy S. DeLoache is a Professor of Psychology at the University of Illinois at Urbana-Champaign. Address correspondence to Judy DeLoache, Department of Psychology, University of Illinois, 603 East Daniel, Champaign, IL 61820.

SYMBOLIC DEVELOPMENT

Becoming a proficient symbolizer is a universal developmental task; full participation in any culture requires mastery of a variety of culturally relevant symbols and symbol systems, in addition to language and symbolic gestures. Children make substantial progress in this task in the first years of life. In Western societies, older infants and toddlers start to learn about pictures and pictorial conventions. Most preschool children are taught the alphabet and numbers, many begin to read, and some even start to do simple arithmetic. Many young children also encounter a variety of less common symbols, such as maps, models, musical notation, and computer icons.

Symbolic development plays a prominent role in many theories of child development, and there is a substantial body of empirical work focusing on the development of particular symbol systems, especially drawing, reading, and mathematical competence.[1] My research addresses the general issue of how very young children first gain insight into novel symbol–referent relations and how they begin to use symbols as a source of information and a basis for reasoning.

In our research, my colleagues and I present young children with a particular symbolic representation—most often a scale model, picture, or map—that provides information needed to solve a problem. Use of the symbol requires (a) some awareness of the relation between symbol and referent, (b) mapping the corresponding elements from one to the other, and (c) drawing an inference about one based on knowledge of the other. The majority of our research has involved scale models. Because young children rarely, if ever, encounter real models in which the symbol maps onto a specific referent, we can use scale mod-