`To appear in:`
`Proceedings of the Seventeenth Annual Conference of the Cognitive Science Society,`
`Hillsdale, N.J., LEA, 1995.`

Developing Object Permanence: A Connectionist Model

Denis Mareschal
Experimental Psychology
Oxford University
South Parks Road
Oxford OX1 3UD
`denis@psy.ox.ac.uk`

Kim Plunkett
Experimental Psychology
Oxford University
South Parks Road
Oxford OX1 3UD
`plunkett@psy.ox.ac.uk`

Paul Harris
Experimental Psychology
Oxford University
South Parks Road
Oxford OX1 3UD
`harris@vax.ox.ac.uk`

Abstract

When tested on surprise or preferential looking tasks, young infants show an understanding that objects continue to exist even though they are no longer directly perceivable. Only later do infants show a similar level of competence when tested on retrieval tasks. Hence, a developmental lag is apparent between infants' knowledge as measured by passive response tasks, and their ability to demonstrate that knowledge in an active retrieval task. We present a connectionist model which learns to track and initiate a motor response towards objects. The model exhibits a capacity to maintain a representation of the object even when it is no longer directly perceptible, and acquires implicit tracking competence before the ability to initiate a manual response to a hidden object. A study with infants confirms the model's prediction concerning improved tracking performance at higher object velocities. It is suggested that the developmental lag is a direct consequence of the need to co-ordinate representations which themselves emerge through learning.

Introduction

This paper presents a connectionist model of the development of object permanence on a task involving visual pursuit. Object permanence is the understanding that objects continue to exist independently of direct perception. It is a central theme in the study of infant cognitive development. Piaget's (e.g., 1952) now classic studies relied on the active search for and manual retrieval of hidden objects to gauge the infant's understanding of object permanence. If a baby reached for a visible object but failed to reach for the object when an occluding screen was lowered in front of the object, Piaget concluded that the infant did not understand that the object continued to exist behind the occluding screen. It was not until 7.5 to 9 months of age that infants succeed at this task.

While Piaget's findings are highly replicable, a different experimental paradigm suggests far more precocious abilities in infants. Studies using preferential looking or surprise as the dependent measures, instead of active manual search, suggest that infants as young as 3.5 months understand that objects continue to exist when hidden (e.g., Baillargeon, 1993; Spelke, 1994). These infants will respond differentially when some property (such as solidity) of a hidden object is violated as compared to when no violation occurs. Hence, a developmental lag is evident between infants' understanding of object permanence as measured by passive response tasks and their ability to demonstrate that knowledge in active retrieval tasks. The lag cannot simply be due to a motor control problem since infants can retrieve a visible object by 4 months (von Hofsten, 1989). Furthermore, preferential looking studies suggest that 5.5 month olds are able to differentiate possible from impossible actions for retrieving a hidden object under some circumstances (Baillargeon, 1993).

The origins of the developmental lag, and what it reflects about the underlying mental representations required for completing both tasks is a key question for current infant research. A number of hypotheses have been advanced to address this question. One suggestion is that the nature of the representations which underlie perception and action in these tasks are radically different (Spelke, Katz, Purcell, Ehrlich & Breinlinger, 1994) and that these develop at different rates. A related suggestion is that the underlying representations for the two tasks are the same, but that the perception and action knowledge domains are encapsulated: there is no transfer of learning from one domain to another (Spelke, 1994). According to this view, the developmental lag simply reflects that fact that infants begin practice with manual retrieval at a later age. Finally, a third suggestion is that the underlying representation of an object develops along a continuum such that the representation required to elicit a perceptual response is simply an early state of the representation required to elicit a retrieval response (Fischer & Bidell, 1991; Munakata, McClelland, Johnson, Siegler, 1994).

We propose that there are 2 distinct factors which contribute to the pattern of results outlined above. First, that hidden objects fail to elicit the same level of response in infants as visible objects is due to the need for stronger, more consolidated representations of objects in the former case than the latter. Hence, we subscribe to the view that the underlying representation of an object develops along a continuum. Second, we assume that the manual retrieval of an object typically involves an integrated response requiring the *coordination* of information about an object's identity and position. In contrast, the predictive visual pursuit of an object need not involve information about an object's identity, only its position (Day & Burnham, 1981). Hence, we subscribe to the view that the task demands imposed by manual retrieval involve exploiting and coordinating distinct representational components while predictive visual pursuit need only refer to the representation of an object's position.

We explain the developmental lag between predictive pursuit of hidden objects and manual retrieval of hidden objects as a consequence of the differential task demands for the two behaviours. Manual retrieval requires the coordination of representations while predictive visual pursuit does not. We suppose that the coordination of representations itself needs to be learnt. Hence tasks requiring the coordination of representations will be developmentally delayed in relation to tasks that do not require this extra level of representational integration. Note that the manual retrieval of displaced hidden objects constitutes the most difficult case for the child from this perspective. Not only must the child coordinate representations of the object's identity and position, but they do so in the absence of direct perceptual cues. In contrast, predictive visual pursuit of a visible object constitutes the easiest case for the child—no coordination of distinct representations is required and direct perceptual input is available to support predictions about the object's trajectory. The predictive visual pursuit of hidden objects and the manual retrieval of visible objects constitute tasks of intermediate difficulty on this perspective. In particular, the manual retrieval of visible objects should be easier than that of hidden objects because the latter require the representations that coordinate object position and identity to become more *strongly* established.

This paper describes a working model of the development of object permanence in the domain of visible and occluded visual pursuit tasks (cf. Bremner, 1985 for a review). We implement a computational model that learns to establish the identity of an object in terms of its distinguishable features, that learns to predict the future position of an object on the basis of its recent trajectory and that learns to initiate a manual retrieval response based on a composite representation of an object's position and identity. The mechanisms brought to bear on the computation of the object's position and identity are quite separate. However, they are exposed to the same input stimulus throughout training and thus have the same opportunity to learn about the relevant characteristics of the environment. The architecture of the model is constrained in such a way that predicting the position of an object takes no account of the object's identity. In contrast, the inititation of a manual retrieval response requires a sensitivity to both position and identity. These architectural constraints are motivated by the findings cited above suggesting the independent representation of spatial and featural information by infants. Our purpose in building the model is to explore the viability of the view that the relatively late emergence of the ability to retrieve displaced hidden objects is due to the combined requirements to develop *strong* internal object representations and to *coordinate* those representations. The model is continuously tested on its ability to predict the next position of visible and hidden objects (predictive visual pursuit) and on its potential to retrieve visible and hidden objects (manual retrieval). We can, therefore, establish a developmental profile of the mastery of these skills in the model and compare this profile to that observed in infants. Furthermore, we can manipulate systematically various features of the model in order to determine their effect on performance. The facility to manipulate characteristics of the model permits us to determine the essential properties that govern its performance and to generate novel behaviours that can be evaluated against the experimental literature or inspire new experiments with infants.

The Model

Figure 1 shows a schematic outline of the model. It consists of a modular architecture. Each functional module is enclosed by a dashed line. Note that some units are shared by two modules and serve as a gateway for information between the modules. In accordance with recent neurological evidence (Ungerlieder & Mishkin, 1982) spatio-temporal information about objects in the world is processed independently of featural information. Information enters the network through a 2-dimensional retina homogeneously covered by feature detectors. It is then funnelled down one pathway which processes the temporal history of the object and another which develops a spatially invariant feature representation of the object (Foldiak, 1991).

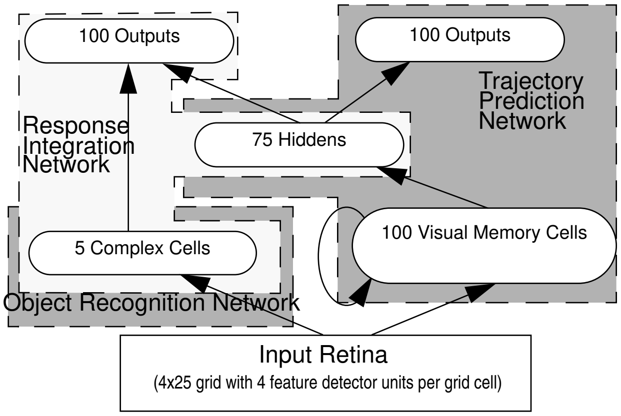


Figure 1: Schematic of modular network architecture.

The retina consists of a 4x25 cell grid. Each cell contains four feature detectors responding to different properties (e.g. light/dark, high/low contrast, hot/cold, soft/hard). If a projected object image overlaps with a grid cell, the cell's feature detectors take on the value +1.0 if the feature is present and -1.0 if the feature is absent (Treisman & Sato, 1990). Cells on which the object image is not projected are quiescent and take on the value 0.0. An occluding screen is also projected on the retina. The cells corresponding to those positions have a constant value of 1.0 and do not encode object features.

The network experiences 4 different objects with correlated features (i.e., {-1 1 -1 1}, {-1 1 1 -1}, {1 -1 1 -1}, {1 -1 -1 1}). All object images are 2x2 grid cells large. For each object presentation, an object moves once back and forth across the retina, either horizontally or vertically. Vertical movements can result in either non-occluding or occluding events while all horizontal movements involve an occluding event. Note the ambiguity when predicting the next position of the object based on a snapshot of its current position. There are four possible next positions for the object: up, down, left, or right. This can only be resolved by learning to attend to the temporal history of the object.

The object recognition module generates a spatially invariant representation of the object by using an unsupervised learning algorithm. That is, it learns to partition the world into consistent feature clusters (to respond similarly to similar objects) without explicit teaching. The feature representation is encoded on a bank of 5 complex cells. These cells are initially randomly connected to all feature detectors. The module develops its representations by using a modified version of the algorithm developed by Foldiak (1991)[1]. This algorithm exploits the fact that an object tends to be temporally contiguous with itself. Thus two successive images will probably be derived from the same object. Learning results in an object (defined by a unique set of features) activating the same complex cell irrespective of its position on the retina.

The trajectory prediction module uses a partially recurrent feed-forward network trained with the backpropagation learning algorithm[2]. At each time step information about the visible position of the object image is extracted from the retina. The retinal grid cells with which the object image overlaps become active (+1.0) while the other cells remain inactive (0.0) (Recall that the trajectory prediction module does not encode feature information about object identity.) These 100 values are mapped one-to-one onto 100 units in the visual memory layer. The network is trained to predict the next instantaneous position of the object. The result is output on a bank of 100 units coding position in the same way as the inputs into the module. The network has a target of +1.0 for those units corresponding to the next object position and 0.0 for all other units.

All units in the visual memory layer have a self-recurrent connection. This gives them the power to process temporal information and generate a representation of the object's spatio-temporal history[3]. The result is a spatial distribution of activation in the form of a comet with a tail that tapers off in the direction from which the object has come. The length and distinctiveness of this tail depend on the velocity of the object. The information in this layer is then forced through a bottle-neck of 75 hidden units. It is here that the network has to generate a more compact, internal re-representation of the object's spatio-temporal history. As there are no direct connections from the input to the output, the network's ability to predict the next position is a direct measure of the reliability of its internal object representation. We suggest that the responses of the trajectory prediction network correspond to the responses observed in infants through the use of (passive) preferential or surprise measures (e.g., Baillargeon, 1993; Spelke, 1994). They are a test of the infant's sensitivity to violations of object position.

The output of the response integration network corresponds to the infant's ability to co-ordinate and use the knowledge it has about object position and object identity. This network is designed to integrate the internal representations generated by other modules (i.e. the feature representation at the complex cell level and spatio-temporal representation in the hidden unit layer) as and when required by a response task. It consists of a single-layered backpropagation network whose task is to output the same next position as the prediction network for two of the objects, and to inhibit any response (all units set to 0.0) for the other two objects. This reflects the fact that infants do not respond (e.g. reach) for all objects. Some objects are desired (e.g. sweet) whereas other are not desired (e.g. sour). Active intentional response necessarily require the processing of featural as well as trajectory information.

Occluded Tracking

The network learns very quickly to predict an object's next position when it is in sight. Moreover, the hidden unit representations that it develops persist even when the object has disappeared and allows the network to keep track of the object even when it is no longer directly perceptible. Figure 2 shows a graphic representation of the network's ability to predict the next position of an occluded object. The left-hand column shows what is projected onto the retina once featural information has been removed. The right-hand column shows the corresponding object position predicted by the trained trajectory network. The rows (from top to bottom) correspond to successive time steps. This network has seen 30,000 presentations of randomly selected objects moving back and forth in random positions and directions at a fixed speed.

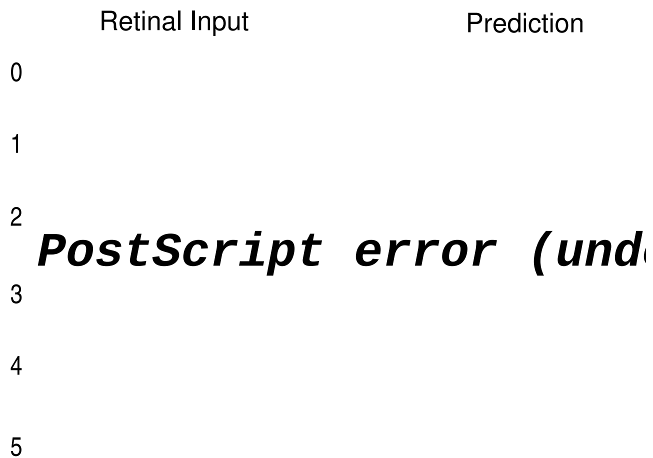
At $t = 0$, the object is about to disappear behind the occluding screen. At all subsequent time steps, the network correctly predicts that the object will have moved over one position. Note especially step 3 for which the direct perceptual information available to the network is exactly the same as at $t = 2$. The network is able to pre-

---

[1] Setting the activations of the loosing units in the competitive phase to a small negative value ($-\beta$) greatly increases the stability of the representations under continued training. We used the following parameter values: $\delta = 0.1$, $\beta = 0.02$, learning rate $\varepsilon = 0.001$, and weight_range $= 0.2$.

[2] All back-propagation networks used the following parameter values:learning rate $\varepsilon = 0.1$ and momentum $\eta = 0.3$.

[3] The recurrent connections were fixed at $\mu = 0.3$.



dict the subsequent reappearance of the object taking account of how long it has been behind the screen. Moreover, as found with infants (Muller & Aslin, 1978), the network's ability to track an occluded object depends on the length of the occluding screen: the longer the screen, the worse the performance.

Developmental lag

The model was designed to examine the developmental lag between an infant's implicit knowledge of object permanence (predictive visual pursuit) and its ability to demonstrate that knowledge with an appropriate response (manual retrieval). Figure 3a shows the network performance (averaged across 10 randomized replications) on both the manual retrieval and visual pursuit tasks when presented with an unoccluded desired object. The reliability of a module is computed as (1 – sum-of-squared-errors across outputs) averaged over the output units and patterns involved in the event. In this case, the network learns very quickly to track and to identify the desired object and to produce an appropriate retrieval response.

When the object is occluded the network's behavior is very different (Figure 3b). Tracking and retrieval responses are initially equally poor. The internal representations are not adequately mature to support any reliable response. At about 5000 epochs they begin to diverge. The reliability of the predictive visual pursuit develops faster than that of the integrated manual retrieval response. Around 20,000 epochs there is a consistent difference between the network performance in the two different modes. The accuracy differential on the two tasks then disappears with further training.

Note that the manual retrieval response which is required for a desired object has exactly the same mode of representation as that for predictive visual pursuit. Moreover, both sets of output units receive exactly the same information from the hidden units about the spatio-temporal history of the object. The only way the functioning of these two modules differs is that the module that drives manual retrieval must integrate information coming from the object recognition module. This indicates that the developmental lag in the network arises from the added task demands of integrating information.

An advantage of modeling is that we can test this hypothesis directly thanks to a manipulation which would not be possible with infants. If the developmental lag is indeed due to the need for an integration of information, then it should disappear when presented with a task that does not require information integration. One possibility is to observe the network's behavior when presented with an undesired object. Undesired objects do not require information integration because it suffices to attend only to the feature representation in order to elicit a proper response, i.e, not to retrieve the object. Once the object has been identified as undesirable, then an inhibitory output can be emitted which does not require any spatio-temporal information. Figure 3c show the network's performance when presented with an undesired object. It learns more quickly to inhibit attempts at retrieval than to track objects. The feature recognition module learns to categorize the different object types very quickly. Hence, it is the need to attend to and integrate information from different sources that produces the lag between reliable predictive visual pursuit and manual retrieval of occluded objects.

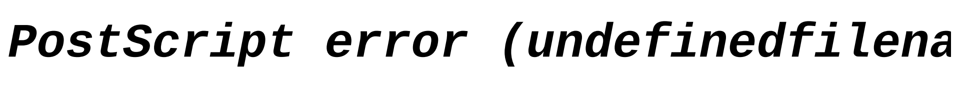


Figure 3: Network performance on predictive visual pursuit and manual retrieval to (a) an unoccluded desired object, (b) an occluded desired object, and (c) an occluded undesired object.

A Test of the Model's Fit

Figure 4a shows the reliability in tracking produced by the network as a function of the velocity of the object image (fast

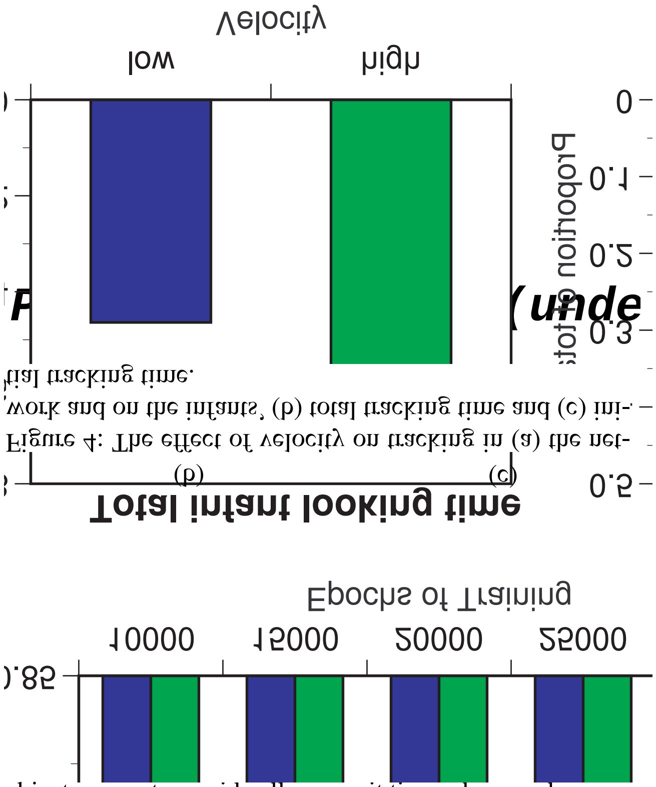


Figure 4: The effect of velocity on tracking in (a) the network and on the infants' (b) total tracking time and (c) initial tracking time.

objects move two grid cells per unit time whereas slow move one grid cell every two time steps). There is consistently better performance at the higher velocity. The greater accuracy arises from the more distinct representations generated in the visual memory layer at higher velocities.

We have found that infants also show superior tracking at higher velocities[4]. In the study, thirty-six 2- to 6-month-old infants sat 0.6m from a viewing theatre and watched an 8° black and white bull's-eye move back and forth across the 1.5 m theatre at either 8 or 12 °/sec. Figure 4b shows the proportion of total tracking time to total visible time that infants spent tracking an object moving across a viewing theatre. Infants showed significantly more tracking in the high velocity condition than the low velocity condition (F(1, 33) = 7.506, p = 0.0098) supporting the predictions of the model.

The total looking time can underestimate the power of a moving object to elicit tracking since subsequent captures and looking times can be artificially reduced due to infant habituation. Since habituation is not implemented in this model, initial infant tracking time may be a better test of the model. Figure 4c shows the proportion of looking time for the initial tracking interval with infants at 0.6m and 1.2m from the target. The velocities at the far distance were 6 °/sec and 8 °/sec. The velocities at the near distance are 8 and 12 °/sec. These correspond to a constant linear velocity difference of 4.2 cm/s at both distances. The object was scaled to subtend 8° at either distance. At both distances there is a significantly longer initial track at higher velocities (F(1, 33)=6.577, p<0.0151) with no significant effect of distance (F(1, 33)=0.291, p<0.5933) or distance by velocity interaction (F(1, 33)=2.548, p<0.12). Again, the main effect of velocity supports the model's predictions.

Discussion

This model suggests that connectionist style learning algorithms are powerful enough to develop perceptually independent representations of objects. These representations allow the network to keep track of an object's properties such as position, velocity, and feature descriptions even when the object is fully occluded. Moreover, there is a gradual emergence of these representations as opposed to an all-or-none acquisition. The representations are not present from the onset and are developed through the interactions of computational-architectural constraints and interactions with the environment.

A critical characteristic of the approach taken in this work is the postulation of a small number of different mechanisms attuned to particular aspects of the environment (cf. Baillargeon, In press). The model assumes the existence of a mechanism designed to compute object identity and a mechanism designed to track object position. Each module learns independently from the same, common experience. The asymmetry in performance on the manual retrieval task and the predictive visual pursuit task is a direct consequence of the requirement that computations delivered by both mechanisms need to be integrated for the former task but not for the latter. It should be noted, however, that the implications of this approach extend beyond the domains of visual pursuit and manual retrieval. In general, *any* task that demands the integration of the computations from distinct modules is likely to be developmentally delayed compared to a task that requires the computations to be delivered from either one of the modules. Of course, the degree of delay observed will depend on the difficulty of the integrative process itself. Thus, in the current model the integration of the computations is particularly difficult for the manual retrieval of hidden objects.

The model also enables to make predictions about infant reactions when objects suffer feature violations. Recall (see Figure 1) that the object recognition network receives direct input from the retina. The complex cells develop spatially invariant object representations from the very start of learning. As these representations become consolidated with training they will tend to persist[5] over time, even when the object is occluded. In other words, the complex cells retain a representation of the object's properties even when

---

[4] The results reported here are part of those obtained during a study designed to test some predictions of the model and to investigate the role of egocentric and allocentric cues in infants' visual pursuit. A full report is in preparation.

[5] The degree of persistence is primarily determined by the parameter δ (see Footnote 1).

the object is out of sight. This information is available to drive a surprise response. Moreover, note that the model predicts precocious behaviour in this domain since the surprise response does not require the integration of computations from distinct representational modules. Knowledge of object properties, such as size, can be driven by computations from a single source. We are currently implementing this extension of the model.

In the future, the model offers further opportunities to investigate the interactions between recognition, visual tracking, and object permanence. Empirical studies have suggested that when a different object reappears from behind the screen, it is the novelty of the object that determines whether infants interrupt their tracking (Goldberg, 1976) and not the change itself. Similarly, this model would suggest that a novel object would disrupt tracking, but only when the change was to an undesired object or one with radically different features. We continue to investigate these interactions both in the model and with infants.

In summary, we suppose that objects are represented and develop in a fragmentary fashion in the child's cognitive system. Different properties of the object (e.g. featural versus spatial-temporal information) are processed in functionally independent modules. The manner in which these properties are brought together depend upon the task demands. The level of object knowledge that the child demonstrates may vary according to the requirements of the task itself.

Acknowledgements

This work was funded in part by the MRC (UK) and FCAR (Quebec).

References

Baillargeon, R. (1993). The object concept revisited: New directions in the investigation of infant's physical knowledge. In: C. E. Granrud (Ed.), *Visual perception and cognition in infancy*, 265-315. London, UK: LEA.

Baillargeon, R. (In press) A model of physical reasoning. In C. Rovee-Collier & L. Lipsitt (Eds.), *Advances in infancy research*, ***9***, Norwood, N.J.: Ablex

Bremner, J. G. (1985). Object tracking and search in infancy: A review of data and a theoretical evaluation, *Developmental Review*, **5**, 371-396.

Day, R. H. & Burnham, D. K. (1981) Infants' perception of shape and color in laterally moving patterns. *Infant Behavior and Development*, **4**, 341–357.

Fischer, K. W. & Bidell, T. (1991). Constraining nativist inferences about cognitive capacities. In: S. Carey & R. Gelman(Eds.), *Epigenesis of mind: Essays on biology and cognition*, 99-126, Hillsdale, NJ: Erlbaum.

Foldiak, P. (1991). Learning invariance in transformational sequences. *Neural Computation*, **3**, 194-200

Goldberg, S. (1976). Visual tracking and existence constancy in 5-month-old infants. *Journal of Experimental Child Psychology*, **22**, 478-491.

Muller, A. A., & Aslin, R. N. (1978). Visual tracking as an index of the object concept. *Infants Behavior and Development*, **1**, 309-319.

Munakata, Y., McClelland, J. L., Johnson, M. N., & Seigler, R. S. (1994). Now you see it now you don't: A gradualistic framework for understanding infant success and failures in object permanence tasks. Technical Report, PDP.CNS.94.2, Carnegie Mellon University, Pittsburg, USA.

Piaget, J. (1952). *The Origins of Intelligence in the Child*. New York: International Universities Press.

Spelke, E. S. (1994). Early knowledge: Six suggestions, *Cognition*, **50**, 431-445.

Spelke, E. S., Katz, G., Purcell, S. E., Ehrlich, S. M. & Breinlinger, K. (1994) Early knowledge of object motion: continuity and inertia. *Cognition*, **51**, 131–176.

Treisman, A. & Sato, S. (1990). Conjunction search revisited. *Journal of Experimental Psychology: Human Perception and Performance*, **16**, 459-478.

Ungerlieder, L. G. Mishkin, M. (1982). Two cortical visual systems. In: D. J. Ingle, M. A. Goodale, & Mansfield (Eds.), *Analysis of visual behavior*. Cambridge, MA: MIT Press.

von Hofsten, C (1989). Transition mechanisms in sensorimotor development. In: A. de Ribaupierre (Ed.), *Transition mechanisms in child development: The longitudinal perspective*, 223-259. Cambridge, UK: Cambridge University Press.