Math Garden: A new educational and scientific instrument
Straatemeier, M.

Citation for published version (APA):
Children's knowledge of the earth: A new methodological and statistical approach

This chapter has been published as:
Abstract

In the field of children’s knowledge of the earth, much debate has concerned the question of whether children’s naive knowledge – that is, their knowledge before they acquire the standard scientific theory – is coherent (i.e., theory-like; Vosniadou & Brewer, 1992) or fragmented (Nobes et al., 2003). We conducted two studies with large samples (\(N = 328\) and \(N = 381\)) using a new paper-and-pencil test, denoted the EARTH (EArth Representation Test for cHildren), to discriminate between these two alternatives. We performed latent class analyses on the responses to the EARTH to test mental models associated with these alternatives. The naïve mental models as formulated by Vosniadou and Brewer (1992) were not supported by the results. The results indicated that children’s knowledge of the earth becomes more consistent as children grow older. These findings support the view that children’s naive knowledge is fragmented.

Introduction

By late childhood, most children, at least in industrialized countries, know that the earth is a sphere that orbits the sun even though it appears to be flat and motionless. But how do they acquire this counterintuitive theory of the earth? Theoretical perspectives on this question differ greatly, and the answer remains unclear. It is, at least, unlikely that children adopt this theory without direct instruction or otherwise culturally transmitted information (Nobes, Martin, & Panagiotaki, 2005). However, children’s own observations and intuitions may also play a role.

The issue of children’s knowledge of the earth boils down to the question of whether children’s naive knowledge – that is, their knowledge before they acquire the standard scientific theory – is coherent (i.e., theory-like) or fragmented. This question of coherence versus fragmentation may be raised more generally in relation to conceptual development. The coherence side of this debate is occupied by the theory theorists who claim that children’s understanding is organized into coherent and consistent theories that structure everyday thinking and are resistant to fluctuations (Wellman & Gelman, 1998). The fragmentation side is occupied by the adherents of the knowledge-in-pieces view who maintain that children’s naive ideas are fragmented. DiSessa (1988) stated that “intuitive physics is a fragmented collection of ideas, loosely connected and reinforcing, having none of the commitment systematicity that one attributes to theories” (p.50). These fragments (i.e., phenomenological primitives: “p-prims”) are small, self-explanatory knowledge structures that are abstracted from experience: they “simply happen.” Both sides continue to pro-
duce empirical support for their respective views (for a review see diSessa, 2006; diSessa, Gillespie, & Esterly, 2004).

Our current aim is to advance this debate in the field of children’s knowledge of the earth by identifying relevant methodological and statistical limitations of previous studies of children’s knowledge of the earth. We demonstrate solutions to these limitations in two new experiments. We first discuss the two competing views of children’s knowledge of the earth. Subsequently, we identify and discuss several methodological and statistical limitations of the current empirical studies in this area. We then present two experiments with a new paper-and-pencil test that we call the EARTH (EArth Representation Test for cHildren).

**Mental models versus fragmented knowledge**

Vosniadou and colleagues are theory theorists in that they hypothesize that children’s naive knowledge of the earth is theory-like, that is, coherent and internally consistent (e.g., Vosniadou & Brewer, 1992). They argued that children construct coherent mental models of the earth; we refer to their view as the mental model account. There are several views concerning mental models in the psychological literature (e.g., Goodwin & Johnson-Laird, 2005; Halford, 1993; Halford & Busby, 2007; Held, Knauff, & Vosgerau, 2006; Holland, Holyoak, Nisbett, & Thagard, 1986; Johnson-Laird, 1983, 2004; Vosniadou, 2002). A mental model is defined as a mental representation that is analogous to the state of affairs the model represents (Johnson-Laird, 1983). The concept of mental models is often used in reference to people’s reasoning (Johnson-Laird, 2004), such as conditional reasoning (Markovits & Barouillet, 2002). It is assumed that people construct one or several mental models when asked to produce deductive inferences. Vosniadou and colleagues (e.g., Vosniadou, Skopeliti, & Ikospentaki, 2004) also argued that mental models are dynamic situation-specific representations formed on the spot for the purpose of answering questions. However, children are assumed to have only one model at a given moment. This model is assumed to have predictive and explanatory power and may mediate in the interpretation of information and in theory revision. Vosniadou (2002) claimed that these mental models play an important role in conceptual development and change. Children’s construction of a mental model is based on their observations and everyday cultural influences. Such observations and cultural influences are, however, subject to the constraints of the underlying conceptual structures (Vosniadou, 2002). In this study, we specifically tested the conceptualization of mental models as proposed by Vosniadou and colleagues. Given their account, children should construct one coherent model on the spot (e.g., during a test) in order to answer the questions consistently. The consistency of responses across a range of questions is then a
measure of the coherence of children’s knowledge.

Vosniadou and colleagues (Vosniadou & Brewer, 1992; Vosniadou et al., 2004) found that 80 to 85% of the children form either initial, synthetic, or scientific mental models of the earth. According to Vosniadou (1994) both the initial and synthetic models are embedded in a naive theory of physics, with the initial models being most strongly influenced by certain “entrenched presuppositions” that apply to physical objects in general. These presuppositions are that the earth is flat (i.e., flatness constraint) and that unsupported objects fall down (i.e., support constraint). The initial models are the rectangular earth model and the disk earth model, in which the earth is a flat object, supported by ground, shaped like a rectangle or a disk. When children are increasingly exposed to culturally accepted information about the scientific view of the earth, they try to assimilate this information into their own naive theoretical framework and, thus, form synthetic models (Samarapungavan, Vosniadou, & Brewer, 1996). The synthetic models are subdivided into three classes: the hollow sphere model, in which the earth is a hollow sphere with people living inside on a flat surface; the dual earth model, in which there are two earths: a spherical earth up in the sky, and a flat earth upon which people live; and the flattened sphere model, in which the earth is round on the sides but flat on the top and bottom, which is where people live. At some point in their development, children adopt the scientific model of a spherical earth with people living all around the world. According to Vosniadou (1994), the acquisition of the scientific model involves a major conceptual reorganization that proceeds through the revision and rejection of children’s presuppositions. The initial and synthetic models discussed above are the most prevalent mental models found in studies with children from Western countries. Although the flatness and support constraints are assumed to be universal, significant cultural variations were found in the manifestation of these presuppositions. For example, Indian children were found to have a model of a flat earth that floats on water, which is consistent with aspects of lay culture (Samarapungavan et al., 1996).

In contrast, Nobes and colleagues claimed that children have no strong presuppositions of flatness and support. Rather, children are initially theoretically neutral (Nobes et al., 2003). According to this view, the development of children’s understanding of the earth comes about through the gradual accumulation of fragments of cultural information about the earth. These fragments come piece by piece and may be mutually inconsistent. Nobes and colleagues claimed that children’s knowledge is fragmented and unsystematic until they acquire the coherent cultural scientific theory of the earth. We refer to this view as the fragmentation account, which resembles the view of diSessa (1988) in that children’s naive knowledge is thought to be fragmented. However, whereas diSessa claimed that fragmented knowledge is phenomenological, Nobes and colleagues emphasized the influence of
culture on children’s knowledge. Children can acquire the correct scientific view relatively early if the cultural information is provided.

Several studies (Nobes et al., 2003, 2005; Siegal, Butterworth, & Newcombe, 2004) produced little evidence for strongly entrenched presuppositions or for the presence of initial and synthetic models given that even the youngest children seemed to possess some scientific knowledge. Rather, these studies supported the idea that children have fragmented and incoherent knowledge of the earth. In addition, the studies showed the importance of culture in children’s knowledge acquisition. For example, Siegal and colleagues (2004) found that Australian children, who are given early instruction in cosmological concepts, have considerably more scientifically correct knowledge than do English children.

Methodological issues

Why are the studies inconsistent, with some results supporting the mental model account (e.g., Vosniadou & Brewer, 1992; Vosniadou et al., 2004) and other results supporting the fragmentation account (e.g., Nobes et al., 2003; Nobes et al., 2005; Siegal et al., 2004)? Below we review the research methods that are used to assess children’s knowledge of the earth. Characteristics of these research methods might explain the various outcomes of the different studies. Based on this review, we argue that the preferred method for investigating children’s knowledge of the earth is a paper-and-pencil test with forced-choice items in which all mental models found in industrialized (Western) cultures are represented.

Generative methods

Vosniadou and colleagues (2004) argued that the best way to investigate children’s knowledge of the earth is to use methods that require children to make productive use of the information to which they have been exposed. For this reason, Vosniadou and colleagues prefer interviews with open-ended and so-called generative questions. They noted that factual questions can be answered by simply repeating culturally transmitted knowledge. In addition, children were asked to produce a drawing or a clay model of the earth. Vosniadou and colleagues (2004) argued that generative questions and the tasks of drawing or making clay models encourage children to make generative use of the scientific information that they have at their disposal, and encourage the on the spot formation of mental models.

Nobes and colleagues (2003, 2005) and Siegal and colleagues (2004) have criticized the drawing method of Vosniadou and colleagues. Children may draw an incorrect picture of the earth simply because they are unable to draw a sphere (Blades & Spencer, 1994; Ingram & Butterworth, 1989) or because they have a bias toward orienting drawings of figures to a vertical or horizontal baseline (Pemberton, 1990). Asking children to make a clay model
of the earth (Samarapungavan et al., 1996; Vosniadou et al., 2004) is open to similar criticism and may also be difficult for children (Siegal et al., 2004). Moreover, pictures and clay models, as well as interview responses, are difficult to score objectively. Although interrater reliability within a research group may be very high (Samarapungavan et al., 1996; Vosniadou & Brewer, 1992), the between-research groups reliability might be low – a well-known phenomenon in the scoring of conservation classifications (Brainerd, 1973).

The interview method used by Vosniadou and colleagues (e.g., Vosniadou & Brewer, 1992; Vosniadou et al., 2004) may also be problematic. During the interview, similar questions (i.e., questions addressing the same issue) were asked to clarify children’s answers and to arrive at a full understanding of the underlying conceptual structures (Vosniadou & Brewer, 1992). This results in a prolonged method of repeated questioning. The everyday conventions concerning conversation do not apply in such situations, and this may confuse young children (Siegal, 1997, 1999; Siegal & Surian, 2004). Another potential problem of the interview method is that children might not be familiar with the words that are used (Siegal, 1997). Both research concerning verbal conservation tasks (Bijstra, van Geert, & Jackson, 1989; Donaldson, 1978; Elbers, 1989; McGarrigle & Donaldson, 1975; Rose & Blank, 1974; Siegal, Waters, & Dinwiddy, 1988) and research concerning eyewitness testimonies of young children (Ceci & Bruck, 1993, 1998; Poole & White, 1991) showed that the above-mentioned problems are a considerable source of error. Both under- and overestimation of children’s knowledge and the consistency of this knowledge may occur. In summary, although the open-ended interview method can be useful during an exploratory phase of the study into children’s knowledge of the earth and the different models they construe, a judgment-only (i.e., forced-choice) test is preferred for further research.

**Forced-choice interviews**

Adherents of the fragmentation account (Nobes et al., 2003; Siegal et al., 2004) have used mainly interviews with forced-choice questions that require children to choose between a “scientific” and an “intuitive” alternative. In addition, children needed to choose the correct shape of the earth from several three-dimensional models. The forced-choice questionnaires elicited more scientific responses than did the open-ended interview method of Vosniadou and Brewer (1992). The forced-choice interviews have the advantages of structure and objective scoring, but the complex social interaction may still confound the results.

Vosniadou and colleagues (2004) raised three points of criticism against the forced-choice questionnaires. First, they argued that responses to the forced-choice questionnaires may be biased toward the scientific view because the choice between a scientific answer
and an intuitive one is limited and because of the lack of justifications of answers. However, empirical evidence with respect to conservation of volume (Kingma, 1984; van der Maas & Molenaar, 1996), the class inclusion task (Chapman & McBride, 1992; Hodkin, 1987; Thomas, 1995; Thomas & Horton, 1997), and the balance scale task (Jansen & van der Maas, 1997, 2002; Siegler, 1981) has demonstrated that, if the construction of the test is based on extensive research on possible alternative strategies, it is possible to design a forced-choice task that does not overestimate children’s knowledge and allows for the detection of alternative strategies.

The second point of critique by Vosniadou and colleagues (2004) is that forced-choice methods inhibit the generation of mental models other than the scientific model because these methods encourage children to reason on the basis of the scientific model. A test that contains all mental models found in previous research carried out in industrialized (Western) countries may overcome this problem. Moreover, as demonstrated below, the hypothesis that open-ended interviews and drawing pictures encourage the formation of mental models can be tested by assessing whether children show more consistent knowledge after being tested with either a drawing task or an open-ended interview.

A final point of criticism of Vosniadou and colleagues (2004) is that children perform better on the forced-choice questionnaires because they only need to recognize scientifically correct information instead of retrieving or constructing this knowledge on their own. The advantage of recall tasks (e.g., open-ended interviews) is the assessment of spontaneous thinking, whereas recognition tasks (e.g., forced-choice tests) have the advantages of a more standardized assessment procedure and objective scoring method. Nonetheless, both responses to a forced-choice task and explanations (answers to open-ended questions) are indirect measurements of the knowledge of interest, and there is no reason to assume that explanations are a better reflection of knowledge than responses to a forced-choice task. Moreover, if mental models exist and play an important role in the development of children’s knowledge of the earth, we should be able to detect them with both recognition and recall tasks.

**Non-verbal test**

Nobes and colleagues (2005) attempted to develop a research method that was more standardized than an open-ended interview and contained a wider choice of mental models than was represented in the forced-choice questionnaires. Children and adults needed to rank 16 pictures according to how well each picture represented the earth. Each picture was a combination of three properties of the earth: shape (sphere, flattened sphere, hollow, or disk), location of people (around or on top), and location of the sky (around or on top).
Using this mainly nonverbal test, Nobes and colleagues (2005) did not find any support for the existence of mental models or presuppositions of flatness and support.

Although the test of Nobes and colleagues (2005) is a promising new approach, it does have several limitations. First, the ranking test must be administered in a time-consuming, one-to-one study situation in which aspects of social-interaction and linguistic ability come into play. Second, the ranking test is complex, especially for young children. The pictures of the ranking task vary along three dimensions. Research has shown that young children have trouble with seriation problems, especially when various dimensions come into play (Siegler, 1998). Therefore, the ranking test of Nobes and colleagues assesses not only knowledge of the earth but also the ability to create correct orderings. Third, children may experience problems ranking pictures that are inconsistent with their mental model and, therefore, are incorrect. For example, if children think that the earth is flat, it is possible that they judge all pictures of differently shaped earths as equally bad representations of the earth. Hence, using the ranking of these pictures as information concerning children’s knowledge is questionable. The final problem concerns the analysis of the responses given that few suitable methods of statistical analysis are available for this new nonstandard test format (i.e., a ranking of 16 cards).

**New test: EARTH**

We suggest using a structured nonverbal forced-choice test that can be administered without one-to-one supervision, so that more children can be tested at the same time and the social interaction is minimal. The training of experimenters and the use of complex coding systems are not required. We constructed a paper-and-pencil test (EARTH) in which the most prevalent models found in earlier studies with samples from Western countries (Vosniadou & Brewer, 1992; Vosniadou et al., 2004) are represented given that these models are relevant for children in The Netherlands. Therefore, children have a rather wide choice of answers, and this should enable us to detect the synthetic models found in previous research. In this way we address the points of criticism by Vosniadou and colleagues (2004) and solve the problems of the open-ended interview, the forced-choice interview, and the ranking test.

**Statistical Issues**

Statistical issues also provide explanations for the various results of the different studies. Below we describe and discuss the data-analytic techniques used in previous research. Subsequently, we introduce a technique, latent class analysis (LCA), that has several important advantages compared with these methods.
Several researchers (Samarapungavan et al., 1996; Vosniadou, 1994; Vosniadou & Brewer, 1992, 1994; Vosniadou et al., 2004) used a methodology that resembles Siegler’s rule assessment methodology (Siegler, 1976, 1981). For each mental model, an expected pattern of responses was formulated, and the degree of correspondence between the expected and obtained response patterns was determined. Vosniadou and Brewer (1992) based these expected response patterns on previous research (e.g., Mali & Howe, 1979; Nussbaum, 1979; Nussbaum & Novak, 1976; Sneider & Polus, 1983). In some of the early studies (e.g., Vosniadou & Brewer, 1992), revisions of the expected response patterns were made after examination of the data, and some deviations of the expected response patterns were tolerated. In later research (Vosniadou et al., 2004) no deviations were allowed. Using this method, many children (80-85%) were classified as having a mental model of the earth.

There are several problems with the rule assessment methodology (Jansen & van der Maas, 1997, 2002; van der Maas & Straatemeier, 2008). First, the assignment of response patterns to mental models takes place on the basis of an arbitrary criterion of a minimum percentage of correspondence between observed and expected responses. Moreover, this criterion can be more or less stringent depending on the number of items used in the test, resulting in different classifications (Van der Maas & Straatemeier, 2008). A second problem is that only models that are formulated beforehand can be detected. Hence, the detection of alternative models is not feasible. Third, the rule assessment methodology does not provide statistical information regarding the goodness of fit of the models (i.e., a quantification of how well the model actually accounts for the data).

Latent class analysis (LCA) solves these problems. Because there are many introductions to LCA (e.g., Clogg, 1995; McCutcheon, 1987; Rindskopf, 1987), we provide only a brief explanation here. LCA belongs to the family of latent structure models. Latent structure models make a distinction between manifest variables (observed responses) and latent variables (the underlying or unobserved variables). Latent structure models differ in the measurement level of the variables. In the case of LCA, the measurement level of both the latent variables and the manifest variables is categorical. LCA divides a sample of response patterns into a limited number of groups, the so-called latent classes (McCutcheon, 1987). LCA can be used when people are assumed to differ qualitatively from each other. The latent classes may represent qualitatively different styles, personalities, attachment patterns, developmental stages, strategies, and so on. Children’s responses to the EARTH items can be classified by means of LCA to find possible underlying mental models. For every latent class, the technique gives a probability of being in that class (unconditional probabilities). Furthermore, every class (or mental model) is characterized by a pattern of conditional probabilities. A conditional probability indicates the probability of giving a
particular response to an item given membership of that class. If the mental model account is correct, we should find a limited number of latent classes with conditional probabilities consistent with the mental models, as formulated by Vosniadou and Brewer (1992). For example, children with a flat earth model should have high probabilities of responding that the earth is shaped like a rectangle or disk, and that people can live only on top of the earth.

An important advantage of LCA is that it offers statistical fit measures that indicate how well a given latent class model accounts for the data. The criterion used to classify children, therefore, is not arbitrary but is amenable to statistical testing. Thus, LCA may falsify a given theoretical account. For example, in the case of children’s knowledge of the earth, if a one-class model fits the data, it seems more plausible that children cannot be categorized into qualitatively different mental models of the earth. A single-class model would suggest that children show only quantitative differences in their knowledge. A second advantage is that the technique can be used to model error processes. Although a child may have a coherent mental model, he or she may still produce answers to items that differ from the answers of other children with the same mental model. One cause of such response variation is response errors. Deviations from the expected response pattern can be modeled in a latent class model because the conditional probabilities for a given answer need not equal 0 or 1. Moreover, LCA can detect clusters of unexpected response patterns: therefore, the mental models do not need to be known beforehand. These unexpected response patterns may suggest unanticipated alternative models. LCA has proven its value in a variety of developmental studies (e.g., Boom, Hoijting, & Kunnen, 2001; Jansen & van der Maas, 1997, 2001, 2002; Raijmakers, Jansen, & van der Maas, 2004; Rindskopf, 1987; van der Maas, 1998).¹

In what follows, we present two experiments – a pilot and a main experiment – in which we studied children’s knowledge of the earth. With these experiments, we aim to address the coherence versus fragmentation debate in the field of children’s knowledge of the earth. As argued, we use a structured paper-and-pencil test (i.e., EARTH) in which the most prevalent mental models found in previous research are represented, and we apply LCA to detect possible underlying mental models.

¹Nobes et al. (2005) applied cluster analysis to the responses to the ranking test so as to detect mental models. An advantage of the use of cluster analysis, in comparison with the rule assessment methodology, is that it uses a less arbitrary criterion to assign children to mental models and does not formulate mental models beforehand. However, cluster analysis does not provide goodness-of-fit statistics like LCA does. Moreover, different clustering methods can generate different solutions (Aldenderfer & Blashfield, 1984).
Pilot experiment

In the pilot experiment, 328 children (4-11 years of age) from two different primary schools in The Netherlands completed the first version of the paper-and-pencil test, the EARTH-1, followed or preceded by a drawing task in which children needed to produce a drawing of the earth with people, trees, the sky, the sun, and the moon. More details about the scoring of the EARTH-1 and the assessment procedure can be found in the Method section of the main experiment.

Below we discuss the main findings of the pilot experiment. We elaborate on several of these results in the discussion of the main experiment. (A full description of the pilot experiment can be found at http://hvandermaas.socsci.uva.nl/Homepage_Han_van_der_Maas/EArth.html.) The results challenge the claim of Vosniadou and colleagues (e.g., Vosniadou & Brewer, 1992, 1994) that young children have strong entrenched presuppositions that the earth is flat and that unsupported objects fall down. Although younger children did sometimes choose pictures of a flat earth or pictures where trees and people were placed only on top of the earth, the distribution of children’s scores on the EARTH-1 showed that the children preferred the scientific answer on other items, given that only a small percentage of children had an overall low score (see Figure 2a below). Furthermore, both the amount of scientifically correct knowledge and the consistency of children’s responses increased with age (see Figure 2c). In addition, a high correlation was found between the amount of children’s scientifically correct knowledge and the consistency of their responses, $r_s = .86$, $p < .001$.

Children’s responses to the EARTH were classified into one of the mental models found in industrialized (or Western) cultures. Only a small percentage of the response patterns could be classified as corresponding to a mental model even when deviations were allowed (see Figure 3a). In addition, the LCA did not detect any of the nonscientific mental models as formulated by Vosniadou and Brewer (1992). Vosniadou and colleagues (2004) assumed that the formation of mental models is encouraged by drawing a picture of the earth. However, no differences in consistency score were found between the experimental conditions. The final issue that was addressed is the issue of recognition versus recall. Children scored significantly higher on the EARTH-1 (a recognition task) than on the drawings (a recall task).

Although the EARTH-1 has several advantages, it also has some shortcomings. First, a few items require revision because spontaneous remarks of the children revealed possible misinterpretations of these items. A second shortcoming of the EARTH-1 is that not all mental models were represented in all items. These shortcomings of the EARTH-1 were addressed in a new test version – the EARTH-2.
Main experiment

A new test version, the EARTH-2, was constructed for the main experiment. The items of the EARTH-1 that posed problems in the pilot experiment were revised or removed, and new items were added. Moreover, we used three-dimensional pictures in the EARTH-2. Care was taken to ensure that the mental models of Vosniadou and Brewer (1992) were used consistently throughout the test. We used an open-ended interview, instead of a drawing, as a generative task. Some children in our sample were interviewed either a half-hour before or a half-hour after completing the EARTH-2.

Method

Participants

We tested 196 boys and 185 girls from three different primary schools in The Netherlands with mainly White students. The parents were asked for written permission for their children’s participation. Children in four different age groups, corresponding to kindergarten and the first three grades, were recruited. The reason for this was that the pilot study showed that children in Grade 4 or above had a considerable amount of scientific knowledge of the earth. Participants were 145 children from kindergarten (4-6 years of age, $M = 5.25, SD = 0.61$); 71 children from Grade 1 (6 or 7 years of age, $M = 6.86, SD = 0.35$); 80 children from Grade 2 (7 or 8 years of age, $M = 7.80, SD = 0.46$); and 85 children from Grade 3 (8 or 9 years of age, $M = 8.92, SD = 0.53$).

Materials

EARTH-2

The EARTH-2 was developed for children 4 years of age or older. The EARTH-2 was presented in a booklet that consisted of 10 pages – 1 page for each item. A sample item was provided first to familiarize the children with the item format and the method for answering. The sample item was followed by 9 test items. Every item consisted of a question about the earth with five or six three-dimensional pictures printed below the question. The pictures represented the models that were found in studies with samples from industrialized (Western) countries: disk earth, hollow earth, dual earth, flattened sphere, no gravity sphere, and the scientific model. The dual earth model was represented in only 2 items, namely, those concerning the shape of the earth. Children were expected to show a mixed preference for flat and spherical pictures on the other items. The other models were represented consistently throughout the test. Some
items included two pictures corresponding to the flat earth model. These pictures were included to distinguish between children who were influenced by the flatness constraint but not by the support constraint. The items of the EARTH-2 concerned the shape of the earth, gravity, and the day/night cycle. An example of a test item (Item 7) is displayed in Figure 1. The questions of the items are shown in Table 1. (The complete EARTH-2 can be viewed at hvandermaas.socsci.uva.nl/Homepage_Han_van_der_Maas/EArth.html.) As in EARTH-1, only the word “earth” (aarde in Dutch) was used. The study of Siegal and colleagues (2004) showed that using the word “earth” leads to the same results as using the word “world” (wereld in Dutch).

Figure 1 Item 7 of the EARTH-2: What happens when you walk along a straight line for a very long time?

For every question, one or two parallel questions addressing the same topic were included. For our data analyses, we used both the original nominal responses to the EARTH-2 and an ordinal scoring of the responses. For this ordinal score, participants received 1 point for choosing the disk model, 2 points for choosing the hollow or dual earth, 3 points for choosing the flattened or no gravity sphere and 4 points for choosing the scientific earth. The average score on the EARTH-2 was calculated, resulting in a minimum score of 1 and a maximum of 4. A 3-point score (0 points for the flat earth; 1 point for the hollow, dual, flattened, and no gravity earth; and 2 points for the scientific model) was also used in this experiment to compare the results of the EARTH-2 with the results of the interview. The correlation between these two scoring methods of the EARTH-2 was high, \( r_s = .99 \) (\( p < .001 \)), indicating that the results are robust to different methods for scoring the EARTH-2 responses.
Table 1 Percentage of responses to the items of the EARTH-2.

<table>
<thead>
<tr>
<th>Question</th>
<th>Flat earth</th>
<th>Hollow</th>
<th>Dual</th>
<th>Flattened</th>
<th>No gravity</th>
<th>Scientific</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. What does the earth look like?</td>
<td>4.5</td>
<td>10.3</td>
<td>2.9</td>
<td>1.3</td>
<td>-</td>
<td>81.0</td>
</tr>
<tr>
<td>2. Which picture shows best where the people live on the earth?</td>
<td>16.1</td>
<td>11.6</td>
<td>-</td>
<td>3.4</td>
<td>24.3</td>
<td>44.6</td>
</tr>
<tr>
<td>3. Which picture shows best where the clouds are?</td>
<td>7.7</td>
<td>8.2</td>
<td>-</td>
<td>1.6</td>
<td>37.5</td>
<td>45.1</td>
</tr>
<tr>
<td>4. Which picture shows best what happens when a giant kicks a ball real hard?</td>
<td>Falls off the earth</td>
<td>8.5</td>
<td>10.1</td>
<td>5.8</td>
<td>9.3</td>
<td>26.5</td>
</tr>
<tr>
<td></td>
<td>Does not fall off the earth</td>
<td>5.8</td>
<td>-</td>
<td>9.3</td>
<td>-</td>
<td>26.5</td>
</tr>
<tr>
<td>5. Which picture shows best where the trees are on the earth?</td>
<td>18.3</td>
<td>9.5</td>
<td>-</td>
<td>4.2</td>
<td>25.5</td>
<td>42.4</td>
</tr>
<tr>
<td>6. Where is the sun at night?</td>
<td>Cloud</td>
<td>15.6</td>
<td>29.6</td>
<td>5.0</td>
<td>-</td>
<td>9.0</td>
</tr>
<tr>
<td></td>
<td>Sundown</td>
<td>5.0</td>
<td>-</td>
<td>-</td>
<td>9.0</td>
<td>40.9</td>
</tr>
<tr>
<td>7. What happens when you walk along a straight line for a very long time?</td>
<td>Falls off the earth</td>
<td>11.3</td>
<td>9.8</td>
<td>5.8</td>
<td>5.5</td>
<td>20.1</td>
</tr>
<tr>
<td></td>
<td>Does not fall off the earth</td>
<td>5.8</td>
<td>-</td>
<td>5.5</td>
<td>-</td>
<td>20.1</td>
</tr>
<tr>
<td>8. Which picture resembles the earth best?</td>
<td>3.2</td>
<td>5.5</td>
<td>5.5</td>
<td>4.7</td>
<td>-</td>
<td>81.0</td>
</tr>
</tbody>
</table>

Note. - means that a mental model was not represented in an item, N = 379
For each model included in the test, a consistency score was calculated as the number of times a model was chosen divided by the number of times the model was represented in the test. This resulted in six consistency scores ranging from 0 to 1, where 0 means a child never chose that model and 1 means a child’s answers were completely consistent with that model. Children were assigned to the model for which they had the highest consistency score. This method resembles the rule assessment methodology of Siegler (1976, 1981).

**Interview**

From every class, 2 boys and 2 girls were randomly selected for the interview, resulting in a total of 68 children. The sample consisted of 28 4- to 6-year-olds (kindergarten), 12 6- or 7-year-olds (Grade 1), 14 7- or 8-year-olds (Grade 2), and 14 8- or 9-year-olds (Grade 3). (The nine interview questions and the aids used during the interview can be viewed at http://hvandermaas.socsci.uva.nl/Homepage_Han_van_der_Maas/EArth.html.) All questions had an open-ended format, except for the second question, in which children needed to choose an earth model from five three-dimensional polystyrene foam models. The questions of the interview were based both on the items of the EARTH-2 and on the questions used by Vosniadou and colleagues (Vosniadou & Brewer, 1992, 1994; Vosniadou et al., 2004). The interviewers could resort to standardized follow-up questions if they judged the children’s answer to be incomplete or unclear.

Children received 0 points for an initial answer, 1 point for a synthetic answer, and 2 points for a scientific answer, and the average score on the questions was calculated. For the flat earth, hollow earth, flattened earth, and scientific models, we calculated consistency scores based on the original responses of the children. A consistency score was the number of times an answer consistent with a model was given divided by the maximum number of possible answers consistent with that model. No consistency score for the dual earth model was calculated. Children with a dual earth model should be characterized by a dual earth answer to Question 2 and a mixed pattern of flat and spherical responses to the other questions. However, none of the children combined the flat and spherical model in response to Question 2. Children were assigned to the model with respect to which they had the highest consistency score.

Five raters, who were also the interviewers, scored the answers of all 68 children. The correlations (Spearman’s rho) between the scores of the different raters ranged from .92 to .98, all significant with $p < .001$. Fleiss’s kappa (Fleiss, 1971) for the model classifications was .76, $p < .05$. For the final score of each child, only the judgment of the rater who interviewed that particular child was used.
Procedure

In the case of the older children (6-9 years of age) the EARTH-2 was administered in a classroom setting. Five experimenters were present during the test session to help the children. After handing out the booklets, one of the experimenters stood in front of the classroom and told the children that she was going to ask some questions about the earth. First, the sample item was used to explain the procedure of marking the answers. The experimenter showed an enlarged version of the sample item and read the question out loud. After this, the children were asked to mark their answers. The same procedure was followed with the other questions. The whole procedure took approximately 10 min. Children in kindergarten (4-6 years of age) were tested outside the classroom in small groups of one or two children per experimenter. This procedure was similar to the classroom procedure described above.

The interview was administered individually. Detailed notes were taken during the interview, and all interviews were audiorecorded. Prior to the actual interview, the children were told that some questions about the earth would be asked. The entire interview procedure lasted approximately 10 min. Of the interviewed children, 65 also completed the EARTH-2. These children were interviewed either a half-hour before \( n = 25 \) or a half-hour after \( n = 40 \) completing the EARTH-2.

Results

EARTH-2

The internal consistency of the EARTH-2, expressed by Cronbach’s \( \alpha \), was .79. The responses of the children to the items of the EARTH-2 are shown in Table 1. On most items, the children preferred the scientific answer. This was most evident for Items 1 and 8, which concern the shape of the earth. In addition to the scientific answer, the children often chose the no gravity answer. Children who chose this answer acknowledged that the earth is a sphere but did not completely understand the concept of gravity. Surprisingly, high percentages of children chose pictures of a flat earth on Items 6 and 9, which concern the day/night cycle. This suggests that children who have a scientific model of the earth do not necessarily fully understand the day/night cycle. Except for Items 6 and 9, there was no clear preference for pictures of a flat earth in the young age groups. Children in the younger age groups preferred the no gravity sphere on most items.

For every child, the mean score on the EARTH-2 was calculated. Figure 2b displays the distribution of these mean scores. Only one child consistently chose pictures of a flat earth, corresponding with a score of 1, and only 8.0% of the children had a score of 2 or less. Thus, even most of the youngest children chose scientific or nearly scientific answers (flattened
earth or no gravity sphere) on at least one item. These results are comparable to the findings of the pilot experiment (see Figure 2a).

To assess the influence of age, school, gender, and condition on the mean score on the EARTH-2, an analysis of covariance (ANCOVA) was performed with age as covariate. All findings below were also analyzed and confirmed with nonparametric tests. Age was a significant predictor of the score on the EARTH-2, \( F(1, 369) = 178.78, p < .001, \eta^2 = .33 \). The solid line in Figure 2d shows that the older children were, the more scientifically correct knowledge they had of the earth. Figure 2c depicts similar results for the pilot experiment. There were no significant differences between the schools and experimental conditions in mean scores on the EARTH-2, so we pooled the data in these groups. However, there was a significant effect of gender: \( F(1, 371) = 15.18, p < .001, \eta^2 = .04 \). This effect is discussed later.

**Figure 2** (a) Distribution of the EARTH-1 scores in Experiment 1. (b) Distribution of the EARTH-2 scores in the main experiment. (c) Mean score and consistency score on the EARTH-1 in the pilot experiment by age group. (d) Mean score and consistency score on the EARTH-2 in the main experiment by age group. (e) Mean score and consistency score on the interview in the main experiment by age group.
Mental models

Every child was assigned to the mental model with respect to which he or she had the highest consistency score on the EARTH-2. The response patterns of only 7.2% of the children were completely consistent with one of the mental models, and the response patterns of only three children (0.8%) were completely consistent with a nonscientific model. Figure 3b shows that the percentage of children with a mental model, as established using the EARTH-2, slowly increased when less strict cutoff scores for consistency were used. When a cutoff score of .8 was used – that is, a single deviation in the answers of the children was allowed – only 18.4% of the children were classified as having one of the mental models. However, most of these children had a scientific model of the earth. Only 1.9% of all children had a nonscientific model when a cutoff score of .8 was used. With a more lenient cutoff consistency score of .6, 44.7% of the children were classified as having one of the mental models and only 9.8 % of the children were classified as having a nonscientific model. The same pattern was also found for the consistency scores on the EARTH-1 in the pilot experiment (see Figure 3a).

An ANCOVA was performed to assess the influence of school, gender, condition, and age on the consistency of children’s answers. These results were again checked using non-parametric tests. The effect of age on the consistency of children’s answers was significant: $F(1, 369) = 138.12, p < .001, \eta^2 = .27$. Figure 2d shows that the consistency was higher for older children. As in the pilot experiment, there was a positive association between
the mean score on the EARTH-2 and the consistency score, $r = .67 \ (p < .001)$. The more knowledge children had, the more consistent were their answers. The ANCOVA revealed no significant differences in consistency between the schools.

**Mental models created on the spot?**

If the formation of mental models was encouraged by answering open-ended questions about the earth (Vosniadou et al., 2004), the response patterns of children who completed the EARTH-2 after being interviewed should show greater consistency with one of the mental models than should those of children who completed the EARTH-2 first. However, the ANCOVA showed no significant differences between the experimental conditions: $F(2, 369) = 1.42, p = .24$.

**Gender differences**

The ANCOVA showed no significant difference in consistency between the boys and girls, $F(1, 369) = 2.74, p = .10$. However, a nonparametric test did show a significant difference, Mann-Whitney $U = 15,500.50, p < .05$. Recall that a gender difference was also found for the mean score on the EARTH-2. The boys ($M = 3.09, SD = 0.63, n = 192$) scored significantly higher on the EARTH-2 than did the girls ($M = 2.83, SD = 0.67, n = 184$), and the boys ($M = 0.64, SD = 0.20$) had a significantly higher consistency score than did the girls ($M = 0.60, SD = 0.18$). These differences cannot be attributed to a difference in age because the boys and girls did not differ significantly in mean age, Mann-Whitney $U = 17,352.00, p = .47$.

**Latent class analysis**

Children’s responses to the EARTH-2 were analyzed using LCA. Table 2 shows the fit measures of the latent class models of the responses to the EARTH-2. The fit of a latent class model can be assessed by the log likelihood ratio statistics ($LR$), which quantifies the discrepancy between the observed and expected frequencies. When the responses to all questions were included in the analysis, none of the latent class models fit the data adequately. In all of the models, the expected frequencies deviated significantly ($p < .05$) from the observed frequencies. The two-class model also had the lowest Bayesian Inform-

---

2 If the data are sparse (many cells in the frequency table have a low frequency), the fit measures do not follow the theoretical chi-square distribution and, therefore, the fit measures cannot be tested by using the theoretical chi-square distribution. The parametric bootstrap method can be used to obtain an empirical distribution of fit measures (Langeheine, Pannekoek, & Van de Pol, 1995; Van der Heijden, ’t Hart, & Dessens, 1997). By resampling data generated using the estimated parameters of the model, bootstrapped fit measures are obtained. Counting the number of bootstrapped fit measures that are larger than the original fit measure results in a bootstrapped $p$ value. Because our data set is relatively small in comparison to the number of possible response patterns, the bootstrapped $p$ value, instead of the $p$ value derived from the theoretical distribution, is reported in this chapter.
Bayesian Information Criterion (BIC) (Schwarz, 1978), which considers the LR in relation to the number of participants and the number of parameters, \( LR + \ln(N) \times \text{parameters} \). A relatively low BIC indicates a relatively parsimonious, well-fitting model. More support for the two-class model came from the BIC weights. Wagenmakers and Farrell (2004) showed that the relative fit of latent class models can be interpreted more readily when using the BIC weights, \( w_i(BIC) \), instead of the raw BIC values. A BIC weight can be interpreted as the probability that a certain latent class model is the best model given the data and the set of candidate models. The BIC weights showed that the two-class model had a far better fit than the other candidate models. The latent classes of this two-class model were comparable to the latent class model found in the pilot experiment. Children in the first class had very mixed response patterns and children in the second latent class had a strong preference for scientific answers, except that the children showed mixed responses to Item 9 in both classes.

Because previous analyses showed that Items 6 and 9 (see Table 1) were difficult for the majority of children, these items probably do not distinguish well between the latent classes. Therefore an LCA was also performed without these items. Again most latent class models fit the data poorly. Only the three-class model fit the data (\( p = .05 \)). We describe this latent class model in more detail below.

### Table 2 Fit measures of latent class models of the EARTH-2

<table>
<thead>
<tr>
<th># classes</th>
<th>df</th>
<th>LR</th>
<th>( p^* (LR) )</th>
<th>BIC</th>
<th>( w_i(BIC) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>All items</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>2,812,461</td>
<td>4446.58</td>
<td>.00</td>
<td>8699.31</td>
<td>.0000</td>
</tr>
<tr>
<td>2</td>
<td>2,812,422</td>
<td>3641.86</td>
<td>.04</td>
<td>8072.48</td>
<td>.9999</td>
</tr>
<tr>
<td>3</td>
<td>2,812,383</td>
<td>3504.93</td>
<td>.04</td>
<td>8131.23</td>
<td>.0000</td>
</tr>
<tr>
<td>4</td>
<td>2,812,344</td>
<td>3370.55</td>
<td>.04</td>
<td>8097.65</td>
<td>.0000</td>
</tr>
<tr>
<td>5</td>
<td>2,812,305</td>
<td>3271.29</td>
<td>.02</td>
<td>8152.56</td>
<td>.0000</td>
</tr>
<tr>
<td>6</td>
<td>2,812,266</td>
<td>3180.32</td>
<td>.01</td>
<td>8209.83</td>
<td>.0000</td>
</tr>
<tr>
<td>Without Items 6 and 9</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>112,469</td>
<td>2665.02</td>
<td>.00</td>
<td>6496.13</td>
<td>.0000</td>
</tr>
<tr>
<td>2</td>
<td>112,438</td>
<td>2074.85</td>
<td>.02</td>
<td>6036.41</td>
<td>.7037</td>
</tr>
<tr>
<td>3</td>
<td>112,407</td>
<td>1946.12</td>
<td>.05</td>
<td>6038.14</td>
<td>.2963</td>
</tr>
<tr>
<td>4</td>
<td>112,376</td>
<td>1845.60</td>
<td>.02</td>
<td>6103.65</td>
<td>.0000</td>
</tr>
<tr>
<td>5</td>
<td>112,345</td>
<td>1782.26</td>
<td>.01</td>
<td>6129.25</td>
<td>.0000</td>
</tr>
<tr>
<td>6</td>
<td>112,314</td>
<td>1717.61</td>
<td>.00</td>
<td>6200.98</td>
<td>.0000</td>
</tr>
</tbody>
</table>

Note. df = degrees of freedom, LR = log likelihood ratio, BIC = Bayesian Information Criterion, \( w_i(BIC) \) = rounded BIC weights, \( N = 376 \)

* bootstrapped \( p \)-value
In Figure 4 the conditional probabilities of the three latent classes are displayed. These conditional probabilities indicate the probability that children chose a particular picture on an item given membership of Class 1, 2, or 3. The first class (33%) was characterized by high conditional probabilities of choosing scientific pictures on all questions. The children in the second class (40%) switched between the no gravity sphere and scientific pictures. The children in the third class (26%) had a mixed response pattern.

Figure 4 Conditional probabilities for Classes 1, 2, and 3, and the proportion of children per latent class by age group (unconditional probabilities).
Figure 4 shows the proportion of children per latent class per age group. The proportion of children belonging to Class 1 was very small among the 4- to 6-year-olds (6%). This proportion increased with age. Of the 8- or 9-year-olds, 75% belonged to Class 1. The proportion of children in Class 2, the scientific/no gravity class, was the highest for the 6- or 7-year-olds (42%) and then decreased with age. The majority of 4- to 6-year-olds (70%) belonged to Class 3 (mixed). The proportion of children belonging to this latent class decreased with age. Only 16% of the 8- or 9-year-olds belonged to this class. None of the response patterns of the latent classes corresponded to the initial or synthetic models that were formulated by Vosniadou and Brewer (1992). Latent class models with more classes also did not contain classes associated with these mental models. LCA was also performed with other questions excluded and with the 4-point scale for the items, but no classes corresponding to the initial and synthetic models were found.

### Interview

The internal consistency of the questionnaire used in the interview, as expressed by Cronbach’s α, was .79. The responses of the children to the interview questions, the response coding, and the expected patterns of responses used for calculating the consistency score for each mental model can be viewed at hvandermaas.socsci.uva.nl/Homepage_Han_van_der_Maas/EArth.html. Many children gave scientific responses to the majority of questions, and many children gave initial and synthetic responses to Questions 6, 8, and 9. The low percentages of scientific answers to Questions 8 and 9, concerning the day/night cycle, are comparable to the answers to the day/night questions of the EARTH-2. An ANCOVA was performed to assess the influence of age, school, gender, and condition on the mean score on the interview. Nonparametric tests were performed to confirm the results. The effect of age on the score on the interview was significant: $F(1, 62) = 38.71, p < .001, \eta^2 = .38$. The solid line in Figure 2e shows that the scores on the interview increased with age. Although this figure shows that the 7- or 8-year-olds had a higher mean score on the interview than did the 8- or 9-year-olds, this difference was not significant (Mann-Whitney $U = 88.00, p = .64$). There were no significant differences in terms of mean score on the interview between the schools, experimental conditions, and gender.

Each child was assigned to a mental model based on his or her highest consistency score on the interview. This classification was more subjective than the classification based on the

---

3 An alternative LCA (excluding Questions 6 and 9) was performed, where we regarded only the aspect of a mental model that was referred to in a question (shape of the earth, gravity, edge). We combined the response alternatives that contained the same manifestation of this aspect, ignoring other aspects of the mental model. Only a consistent scientific model was found in the resulting latent class model. No classes corresponding to the initial and synthetic models were detected.
EARTH-2.4 Decisions concerning which answers were in concordance with which mental models needed to be made afterward because no complete list of possible responses could be formulated beforehand. Figure 3b depicts the percentage of children with a mental model based on the interview against cutoff scores for the highest consistency score. The results are comparable to those of the EARTH-2. The response patterns of only 4.4% of the children were completely consistent with a mental model, which was the scientific model for all cases. When the cutoff score was set at .8, 11.8% of the children were classified as having a mental model, with only 2.9% of all children having a nonscientific model (i.e., flat earth model). With a lenient cutoff score of .6, 51.5% of the children could be classified, with only 11.7% of all children having a nonscientific model (i.e., 8.8% flat earth model and 2.9% flattened earth model). These percentages are much lower than those reported by Vosniadou and colleagues in their studies (Vosniadou & Brewer, 1992; Vosniadou et al., 2004).

An ANCOVA was performed with the consistency score on the interview as dependent variable, age as covariate, and school, condition, and gender as independent variables. The effect of age on the consistency score on the interview was significant: $F(1, 62) = 20.25$, $p < .001$, $\eta^2 = .25$. Figure 2e depicts the highest consistency score by age group. The consistency score displayed the same age trend as did the mean score. The older the children were, the more consistent were their answers. Figure 2e shows that the 7- or 8-year-olds had a higher consistency score on the interview than did the 8- or 9-year-olds. However, this difference was not significant, Mann-Whitney $U = 86.50$, $p = .60$). There were no significant differences between the schools, experimental conditions, and gender in their consistency scores.

EARTH versus interview; Recognition versus recall

The correlation between the scores on the EARTH-2 and the interview, which can be seen as a measure of validity, was .55 (Spearman’s rho, $p < .001$). Only 20 children (30.8%) were classified as having a mental model based on both methods. All but 2 of these children were classified into the scientific model based on both methods. All of the other children (69.2%) were classified as nonclassifiable with at least one method.

To further compare the scores on the two research methods, a parametric test ($t$ test) and a non-parametric test (Wilcoxon Signed Rank Test) for two paired samples were performed. A 3-point score (initial = 0, synthetic = 1, and scientific = 2) for the EARTH-2 was used to produce comparable scores between the two methods. There was no significant difference in mean score between the research methods, $t(64) = -1.63$, $p = .11$.

---

*To test the robustness of our classification method, we also compared it with two other more liberal methods. These classification methods led to very similar results.*
Wilcoxon Z = -1.42, \( p = .16 \). Similarly, there was no significant difference in the consistency score of the two test methods, \( t(64) = 0.43, \ p = .67 \), Wilcoxon Z = -0.03, \( p = .98 \). So, the children scored just as high and consistently on a recall task (the interview) as on a recognition task (the EARTH-2).

**Discussion**

The aim of this chapter was to shed some light on the nature and development of children’s knowledge of the earth. Specifically, we wanted to test the coherence and fragmentation view of this knowledge. A new nonverbal paper-and-pencil test, the EARTH, was developed to investigate children’s knowledge of the earth. The problems that may characterize an interview situation, relating to the complex social interaction between experimenter and child, do not apply to the EARTH. Care was taken to address the criticism of forced-choice questionnaires by Vosniadou and colleagues (2004). The most important mental models, as found by Vosniadou and Brewer (1992), were represented in the test, thereby giving the children a wide range of choice. Moreover, the EARTH made it possible to test a large sample of children, allowing us to analyze the data with the statistically more advanced method of LCA. With LCA, we showed that, except for the scientific model, no mental models were evident in the data.

The results of both the pilot experiment and our main experiment do not favor the mental model account. First, no support was found for the existence of strong entrenched presuppositions of flatness and support. Both experiments showed that some children, particularly in the young age groups, chose pictures that represented flat or synthetic mental models, but the percentages of children who chose these answers varied considerably between questions. The main experiment showed that the 4- to 6-year-olds often chose pictures that represented the no gravity sphere, which was used more consistently throughout the EARTH-2 than in the EARTH-1. The mean scores on the EARTH-1 and EARTH-2 (with 1 representing the consistent choice for pictures representing a flat earth and 4 representing the consistent choice for pictures representing the scientifically correct earth) showed that only small percentages of children had a score of 2 or less. Thus, the percentages of children with presuppositions of flatness and support were very small. Moreover, the responses of these children were often not completely consistent with the presuppositions. These results indicate that some children may entertain the assumption that the earth is flat and that unsupported objects fall down, but these presuppositions are not as universal or as strong as Vosniadou and Brewer (1992) claimed. At any rate, the presuppositions are not strong enough to constrain children’s naive knowledge in such a way that children construct ini-
tial or synthetic models. It seems more likely that these presuppositions are fragments of knowledge that coexist with fragments of scientific knowledge.

Second, our results are inconsistent with the claim of the mental model account that children form coherent mental models of the earth. Using the rule assessment methodology, only a minority of the responses to the EARTH-2 and the interview were classified into one of the mental models even when a lenient cutoff score was used. Classifications per age group showed that even with a lenient cutoff score of .6, many 4- to 6-year-olds were not classified and the percentage of unclassifiable children decreased with age. The same results were found with the EARTH-1 in the pilot experiment. The percentages of classified responses are much lower than the 80 to 85% that Vosniadou and colleagues found (e.g., Vosniadou & Brewer, 1992; Vosniadou et al., 2004). However, they are similar to the 35.8% reported by Nobes and colleagues (2005).

Why are the percentages of children who were classified into a mental model based on either our interview or both the EARTH-1 and EARTH-2 much lower than the percentage of classified children in the study by Vosniadou and colleagues (2004)? A possible explanation is that Vosniadou and colleagues based the classification on four questions, whereas eight questions (EARTH-1) or nine questions (EARTH-2 and interview) were used in the current experiments. Our classification criterion, therefore, was more stringent. With respect to the responses to the EARTH-1 and EARTH-2, therefore, the same issue was also investigated with the statistically more advanced LCA, which allows for errors by permitting the conditional probabilities to deviate from 1 and 0. In spite of fitting several latent class models, varying the item sets, and using different scoring methods for the items, we found no clear-cut fitting model for our data. Moreover, the response patterns of the latent classes of the latent class models that did fit the data did not correspond to the initial and synthetic models of Vosniadou and Brewer (1992). Only the scientific model could be identified in the pattern of conditional probabilities in one of the classes. In the main experiment, a latent class was found, with children alternating between the scientific and no gravity responses. Alternating between questions seems more in accordance with the fragmentation account given that it is plausible that these children have both scientific and nonscientific fragments, which are not completely consistent, because the concept of gravity is not completely understood.

Although Vosniadou and colleagues (2004) claimed that children’s mental models do not need to be stable and permanent, children are expected to answer questions consistently during a test administration. The reason for this is that mental models are also assumed to have predictive and explanatory power and are assumed to play an important role in conceptual development and change (Vosniadou, 2002). Not only did we fail to find high
consistencies in children’s responses, we also failed to find support for the dynamic nature of the models. Children did not have more consistent response patterns on the EARTH after producing a drawing (pilot experiment) or answering open-ended questions (main experiment) compared with children who completed the EARTH first.

The results of both experiments also indicated that the consistency of children’s knowledge of the earth increases with age. Furthermore, a strong positive association was found between the amount of scientifically correct knowledge children have and the consistency of this knowledge. Hence, children’s knowledge of the earth comes in fragments. Gradually, scientifically correct fragments may replace naive fragments. The remaining question concerns the nature and content of the fragments of knowledge.

Our findings do not provide direct evidence for the correctness of the fragmentation account. Accepting the absence of mental models in support for the fragmentation account is tantamount to accepting the null hypothesis. To obtain proper support for the fragmentation account, positive predictions should also be formulated and confirmed. Examples of such predictions concern the content of the possible fragments of knowledge and the order in which the fragments are acquired. It may be possible to bridge the gap between the two approaches by adopting a different conceptualization of the mental model account. In both approaches, the end state is the scientific view, which clearly fits the definition of a mental model. Our results cast serious doubt on the mental model account, which assumes that children develop this scientific model by first adopting a series of initial and synthetic theory-like models. But other accounts of mental models (e.g., Halford & Busby, 2007; Markovits & Barrouillet, 2002) may provide a better account of the fragmented knowledge of children who have yet to acquire the scientific view.

In the main experiment, the EARTH-2 data were compared with the interview data, which consisted mainly of answers to open-ended generative questions. No differences were found between the two research methods in the level of observed knowledge of the earth. Vosniadou and colleagues (2004) asserted that a forced-choice method may inhibit the generation of internal models. However, there was no difference in consistency scores found between the two test methods, and for both methods the percentage of children who were classified into one of the mental models was low, as was described above. The comparable results of the EARTH-2 and the interview suggest that a forced-choice test can be just as informative as an interview with open-ended questions. They also suggest that the claim that a forced-choice task is easier and generates false positive errors (because it is a recognition task) does not apply to the EARTH. This is probably the result of the format of the EARTH, in which each item offers the children an extensive choice of responses.

However, in the pilot experiment, different results were found for the drawing task and
the EARTH-1, with children scoring lower on the drawing task than on the EARTH-1. That is why we take a closer look at our drawing data here. Table 3 shows the percentages of mental models found with the drawing task. These percentages are comparable to the results found in the studies by Vosniadou and colleagues (e.g., Vosniadou & Brewer, 1992) given that large proportions of children drew a picture of a flat or synthetic earth. Most 4- or 5-year-olds (80.0%) and 5- or 6-year-olds (50.0%) drew a picture of a flat earth. The percentages of flat earth drawings decreased with age, whereas the percentages of drawings of a no gravity earth and a scientific earth increased with age, \( \chi^2(30) = 107.10, p < .001 \). However, the percentages of older children with a correct scientific picture of the earth were relatively low.

Table 3 Percentages of mental models based on the drawing task

<table>
<thead>
<tr>
<th>Age group (years)</th>
<th>Initial</th>
<th>Flat</th>
<th>Hollow</th>
<th>Dual</th>
<th>Flattened</th>
<th>No gravity</th>
<th>Scientific</th>
</tr>
</thead>
<tbody>
<tr>
<td>4-5 year</td>
<td>10.0</td>
<td>80.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>10.0</td>
</tr>
<tr>
<td>5-6 year</td>
<td>11.0</td>
<td>50.0</td>
<td>5.6</td>
<td>16.7</td>
<td>0.0</td>
<td>11.1</td>
<td>5.6</td>
</tr>
<tr>
<td>6-7 year</td>
<td>0.0</td>
<td>34.2</td>
<td>7.9</td>
<td>21.1</td>
<td>0.0</td>
<td>7.9</td>
<td>28.9</td>
</tr>
<tr>
<td>7-8 year</td>
<td>0.0</td>
<td>16.2</td>
<td>5.4</td>
<td>40.5</td>
<td>0.0</td>
<td>13.5</td>
<td>24.3</td>
</tr>
<tr>
<td>8-9 year</td>
<td>5.0</td>
<td>0.0</td>
<td>0.0</td>
<td>30.0</td>
<td>0.0</td>
<td>20.0</td>
<td>45.0</td>
</tr>
<tr>
<td>9-10 year</td>
<td>0.0</td>
<td>7.8</td>
<td>2.0</td>
<td>25.5</td>
<td>0.0</td>
<td>19.6</td>
<td>45.1</td>
</tr>
<tr>
<td>10-11 year</td>
<td>0.0</td>
<td>2.0</td>
<td>6.1</td>
<td>18.4</td>
<td>0.0</td>
<td>42.9</td>
<td>30.6</td>
</tr>
<tr>
<td>Total</td>
<td>1.8</td>
<td>18.4</td>
<td>4.5</td>
<td>24.2</td>
<td>0.0</td>
<td>20.2</td>
<td>30.9</td>
</tr>
</tbody>
</table>

*Note.* The “initial” category contains drawings with nondistinguishable elements. Drawings in which the required elements were drawn separately instead of drawn in one coherent picture were excluded. \( N = 223 \)

The question then arises whether the EARTH-1 is biased toward scientific responses or whether the drawing task is biased toward nonscientific responses. Nobes and Panagiotaki (2007) carried out a study in which adults were asked to draw a picture of the earth. They found that fewer than half of the drawings could be classified as the scientifically correct picture of the earth. In addition, 15% of the drawings were comparable to children’s naive drawings (i.e., flat earth, hollow earth, or dual earth) as observed by Vosniadou and colleagues. A follow-up interview revealed that the adults found the drawing task to be ambiguous and difficult. Therefore, we conclude that children scored higher on the EARTH-1 not because the EARTH-1 triggers the scientific model but rather because the drawing task poses problems for the children. The comparable results of the EARTH-2 and the interview in the main experiment again confirmed the notion that the low scores on the drawing task may be attributed to the inability of children to draw a correct picture, or to a lack of under-
standing of the drawing task, rather than to a lack of knowledge of the earth.

A question that remains is why our interview method resulted in findings different from those of the interview method of Vosniadou and colleagues (e.g., Vosniadou and Brewer, 1992; Vosniadou et al., 2004). With our method, we found that even young children had a considerable amount of scientifically correct knowledge and that only a small percentage of children had a consistent nonscientific model. In contrast, Vosniadou and colleagues found that most young children had a flat or synthetic mental model. One plausible explanation lies in the methodological differences between the two interview methods. In our interview method, children were asked to choose a three-dimensional model, whereas in Vosniadou and colleagues’ interview method, children needed to draw a picture or make a clay model of the earth. Another possible explanation is that different cultural or educational influences led to different performances of the Dutch children in these experiments compared with the American and Greek children in the studies by Vosniadou and colleagues. This explanation seems less plausible given that the data from the drawing task in the pilot experiment showed results comparable to those in the studies with the drawing task of Vosniadou and colleagues (e.g., Vosniadou and Brewer, 1992).

However, the influence of culture received only limited attention in our experiments because our main focus was on addressing the statistical and methodological limitations of previous research. In the pilot experiment, children from a mainly White school and from a multicultural school participated in the experiments. Although no differences were found between the two schools, no analyses on the individual level were performed. In the main experiment, data concerning children’s specific cultural background were gathered but only a very small sample had a non-Dutch background. The influence of culture, therefore, should be addressed in more detail in future research. As has already been done with interviews (e.g., Diakidoy, Vosniadou, & Hawks, 1997; Nobes et al., 2003; Samarapungavan et al., 1996; Siegal et al., 2004), it would be interesting to conduct a cross-national or cross-cultural study using the EARTH.

In future work, researchers are also advised to investigate children’s knowledge of the day/night cycle in more detail because the EARTH included only two items relating to this issue. Many interview studies (e.g., Diakidoy et al., 1997; Siegal et al., 2004; Vosniadou & Brewer, 1994; Vosniadou et al., 2004) have investigated children’s understanding of the day/night cycle. The findings of these studies may be used to develop structured forced-choice tests.

To conclude, we believe that our methodological and statistical approach constitutes an improvement over existing methods in the empirical comparison of the coherence versus fragmentation view. We do not assert that interviews and drawings should not be used.
These methods are valuable in exploratory studies to collect possible responses and to form initial hypotheses on the kinds of knowledge structures that children use. For follow-up research, however, we believe that the EARTH’s practical and methodological advantages make it preferable to other research methods. Our approach can also be applied to advance the coherence versus fragmentation debate in other domains of conceptual development such as children’s knowledge of the concepts of force, heat, and speciation (e.g., diSessa, et al., 2004; Ioannides & Vosniadou, 2002; Samarapungavan & Wiers, 1997; Vosniadou, 1994).