Title: HOW CHILDREN AND ADULTS PRODUCE AND PERCEIVE UNCERTAINTY IN AUDIOVISUAL SPEECH

Authors: Emiel Krahmer and Marc Swerts

Affiliation: Communication and Cognition, Tilburg University

Running head: Uncertainty in audiovisual speech of children and adults

Full address: Emiel Krahmer
Communication and Cognition
Faculty of Arts
Tilburg University
P.O.Box 90153
NL-5000 LE Tilburg
The Netherlands
e-mail: E.J.Krahmer@uvt.nl
phone: +31 13 4663070
fax: +31 13 4663110

Abstract

We describe two experiments on signaling and detecting uncertainty in audiovisual speech by adults and children. In the first study, utterances from adult speakers and child speakers (aged 7-8) were elicited and annotated with a set of six audiovisual features. It was found that when adult speakers are uncertain they are more likely to produce fillers, delays, high intonation, eyebrow movements, and 'funny faces'. The basic picture for the child speakers is somewhat similar, in that the presence of certain audiovisual cues correlates with uncertainty, but the differences are relatively small and less often significant. In the second study both adult and child judges watched responses from adult and child speakers selected from the first study to find out whether they were able to correctly estimate a speakers' level of uncertainty. It was found that both child and adult judges give more accurate scores for answers from adult speakers than from child speakers and that child judges overall provide less accurate scores than adult judges.

Keywords: question answering, audiovisual speech, children, uncertainty, Feeling of Knowing, Feeling of Another's Knowing, speech production, speech perception

INTRODUCTION

For most people, answering factual questions (What is the color of peanut butter? Who wrote Faust?) has been an almost daily activity ever since they entered elementary school. Some of the questions are easy to answer, while others are more difficult, requiring a longer memory search which may or may not be successful. As a result, a person will typically not be able to answer all questions (although sometimes it may feel like the answer lies on the tip of the tongue), and in those cases where an answer is given, it will be associated with a varying degree of uncertainty.

It has been suggested that people convey this degree of uncertainty to the questioner as a kind of self-presentation; by answering in an 'uncertain manner' a speaker may save face if the answer turns out to be incorrect later on. Smith and Clark (1993) have studied the way speakers signal uncertainty in factual question-answering situations, using the *Feeling of Knowing* (FOK) paradigm originally due to Hart (1965). They found that when speakers are uncertain about the correctness of their answer, they may indicate this using a variety of prosodic cues including fillers, longer delays, and rising intonation, and using linguistic hedges such as 'I guess' or 'I think'. This finding suggests that question-answering at least involves two components, one involving the actual memory *search* and another meta-cognitive one which *monitors* the search process (e.g., Koriat 1993, Nelson & Narens 1990; see Nelson 1993 for an overview). A natural follow-up question would be whether a speaker's signalling of uncertainty is communicative, in that listeners are sensitive to such cues. This question was addressed by Brennan and Williams (1995), who focussed on the perception of uncertainty using a variant of the FOK paradigm referred to as the *Feeling of Another's Knowing* (FOAK). Their experiment revealed that the uncertainty cues that were described in Smith and Clark's work indeed have communicative relevance, since listeners use them to make adequate estimates of the certainty or uncertainty of the speaker.

Smith and Clark (1993) and Brennan and Williams (1995), in their respective studies on the production and perception of uncertainty, largely focus on verbal prosody, i.e., suprasegmental features such as intonation (speech melody), tempo, and pausing

that are encoded in the speech signal itself (Cruttenden, 1986; Ladd 1996). However, these cues offer only a limited representation of face-to-face interaction, in which facial expressions and gestures are important ingredients as well, supporting the information a speaker wants to convey and helping to structure the interaction (see e.g., Barkhuysen et al. 2005a, Clark & Krych 2004, Ekman 1979 among many others). To account for this richness of non-verbal communication, various researchers suggested broadening the definition of prosody to also include visual features, such as facial expressions, arm and body gestures and pointing (e.g., Barkhuysen et al. 2005b, Granström et al. 1999, Keating et al. 2003, Krahmer & Swerts 2004, Munhall et al 2004, Srinivasan & Massaro 2003). Interestingly, it appears that speakers also signal their level of uncertainty visually, and that such combined audiovisual signaling leads to a more accurate perception of uncertainty (Swerts & Krahmer 2005).

The aforementioned studies by Smith and Clark (1993), Brennan and Williams (1995) and Swerts and Krahmer (2005) are only concerned with how *adults* answer questions. While it is obvious that the *kind* of questions asked differ for various age groups (e.g., the peanut butter question is more typical for younger and the Faust for older speakers), it is unclear whether the *process* of question-answering changes over time. It is conceivable that children signal and detect uncertainty differently from adults, either because (1) their meta-cognitive understanding of (un)certainty may be less well-developed than that of adults or (2) because their verbal and non-verbal signalling of (un)certainty during answering is different from that of adult speakers.

Children's grasp of notions like certainty and uncertainty in the context of meta-cognitive development has been the subject of a number of studies in the field of *theory of mind* development. The "theory of mind" describes the ability to reason meta-cognitively about such intentional states as 'beliefs' and 'knowledge', both in one's own mind and in that of others. Even though young children quickly attain an elementary theory of mind, it is generally assumed that meta-memory does not properly develop before elementary school (e.g., Bartsch & Wellman 1995, Flavell 1999, Kreutzer et al 1975, Schneider 2001, Schneider & Lockl 202, Wellmann et al. 2001).

Various studies have addressed the question of when children develop the ability to adequately judge their own uncertainty using the Feeling of Knowing paradigm (e.g., Brown & Lawton 1977, Cultice, Somerville & Wellman 1983, De Loache & Brown 1984, Wellman 1977, Butterfield et al. 1988, Lockl & Schneider 2002). The recent findings of Lockl and Schneider (2002) suggest that there is little change in judgment accuracy between the 7th and the 10th year, which confirms the trends of the earlier studies. Whether the children in this age category are as accurate as adults in judging their own Feeling of Knowing is still essentially an open question.

As far as we know, studies with children that directly address Feeling of Another's Knowing have not been conducted yet, but a number of studies have addressed the perception of uncertainty by children from slightly different angles such as (word)learning. For efficient learning, it is essential to know whether an "informant" is reliable or not, since it is obviously better to learn from a reliable and accurate source than from an unreliable or inaccurate one. Children are generally competent at identifying accurate and inaccurate informants when they are 3-4 years old (e.g., Bisanz et al. 1975, Koenig et al. 2004, Robinson et al. 1995). Speaker certainty is an important cue for learning, and it has been shown that children have learned the difference between expressions of speaker certainty and expressions of speaker uncertainty by the time they are 4 years old as well (Moore et al. 1989, 1990, Matsui et al. 2003). Similarly, Sabbagh and Baldwin (2001) show that children of 3 and 4 years old are capable of understanding a speaker's confidence about word-referent links (which are believed to play an essential role for learning the meaning of new words), provided that such attitudes are verbally expressed.

Children's production and perception of uncertainty in audiovisual speech has not been studied directly before, although various researchers have addressed the development of prosody, facial expressions and gestures in relation to other meta-cognitive aspects. In general, research on prosodic capabilities of school age children does not offer a clear picture. While it appears that the basics are acquired early on (in part even before children actually learn to speak), general prosodic (as well as segmen-

tal) production differences between children and adults remain for some time (e.g., timing of consonant clusters and voice onset time; Pena-Brooks & Hegde 2000). Various studies suggest that the relative frequency of fillers in child speech increases with age (MacWhinney & Osser 1977, Narayanan & Ptamiano 2002, Esposito et al. 2004, Montanari et al. 2004), but if and when children produce fillers in a functional way and with the same frequency as adults is an open issue. Facial expressions and gestures are known to be important for children in a conversational setting (see e.g., Cassell 2004), and there is evidence that children, even more than adults, communicate more effectively when they have access to visual cues besides auditory ones (Doherty-Sneddon & Kent 1996). The development of facial expressions and gestures by children has been studied extensively (see e.g., Schmidt & Cohn 2002, McNeill 1993), and here a similar trend can be observed. Basic gestures and facial expressions are learned early on, but differences between children and adults remain for some time. What is particularly relevant in the current context is that gestures which reflect meta-cognitive aspects are the latest to develop. According to McNeill (1993), meta-cognitive gestures are only acquired at the age of 7.

There appears to be no consensus about the perceptual prosodic capabilities of school-age children. Various researchers have addressed the use of prosodic cues for the comprehension of sentences and phrases, some arguing that school-aged children, even up to 10 years old, do not interpret subtle prosodic cues as effectively as adults do (Cruttenden 1974, 1985, Vogel & Raimy 2002), while others maintain that they do (Beach et al. 1996, Leuckefeld et al. 2003). On a more meta-linguistic level, a number of studies have been devoted to children's usage of prosody for detecting irony and sarcasm, again with equivocal results, some researchers finding that young school children do not use intonation for this purpose (e.g., Ackerman 1983), while others suggest they do (e.g., Keenan & Quigley 1999). The extent to which children use subtle visual cues during communication in a functional way is still largely unexplored.

It is not straightforward to judge and compare the various observations mentioned above (due to differences in the operationalization of uncertainty, the experimental

method, the age and culture of the participants, etc.), but some general trends may be observed. In general, it seems that most children are able to distinguish between certainty and uncertainty (when explicitly marked) by the age of 4, and that they are able to quantify their own uncertainty (in terms of Feeling of Knowing) by the age of 7. When looking at non-verbal cues to uncertainty, the picture is more complicated. Little is known about audiovisual signaling and detection of uncertainty by children, and what is known about the use of audiovisual cues in relation to other meta-cognitive phenomena reveals a murky picture.

To gain better insight in the differences and similarities between child and adult signaling and detection of uncertainty during question-answering, we take an experimental approach, performing Feeling of Knowing and Feeling of Another's Knowing studies, both with adults and with school-age children as participants. We opted for 2nd graders (7-8 years old), since this seems to be the youngest age group that according to the literature mentioned above should be capable of making adequate Feeling of Knowing judgements and has developed a basic repertoire of non-verbal cues, although perhaps not yet on an adult level.

We first focus on the *production of uncertainty*, describing two Feeling of Knowing experiments, one with adults and one with child participants. The emphasis in the production experiment is on the different audiovisual cues used by these two groups of participants, drawing on the earlier annotations of Smith and Clark (1993), Brennan and Williams (1995) and Swerts and Krahmer (2005). We hypothesize that children's Feeling of Knowing judgements are similar to those of adults (extrapolating the findings of e.g., Lockl & Schneider 2002), but that children use less audiovisual cues for uncertainty and that they use them in a less systematic way than adults do (given that meta-cognitive cues are among the latest to develop, cf. McNeill 1993). Then, we discuss an experiment focussing on the *perception of uncertainty*, using data collected in the two production experiments as stimuli in a series of perception experiments where children and adults act *both* as speakers *and* as judges. We hypothesize that it is more difficult to judge the uncertainty of child speakers than of adult speakers (assuming

that children in this group indeed signal uncertainty to a lesser extent than adults do), but that adult and child judges make similar estimates of the uncertainty of others, given that, according to the literature, children can distinguish expressions of speaker certainty and uncertainty at the age of 4 (e.g., Moore et al. 1989, 1990).

EXPERIMENT I: SIGNALING UNCERTAINTY

Method

Participants Twenty adults and twenty-one children participated as speakers. The adults (11 males, 9 females) were colleagues and students from Tilburg University, between 20 and 50 years old. They did not object to the usage of their recorded data for research purposes. The children (9 boys, 12 girls) were in 2nd grade (group 4 in the Dutch school system) of 't Schrijverke ("the little writer"), an elementary school in Goirle (a small town adjacent to Tilburg). They were all between 7 and 8 years old. Parents of the children were informed of the experiment in advance, and only those children participated whose parents returned a written statement that they did not object to their child's participation nor to any usage of the recorded material for research purposes.

Stimuli Following Smith and Clark (1993), we used Hart's (1965) method to collect both certain and uncertain speaker utterances from adults and children by asking them a series of factual questions (40 for adults, 30 for children). For adults, the questions came from two sources, where we first selected questions with a one-word answer (e.g. Who wrote Faust? What is the capital of Switzerland?) from a Dutch version of the "Wechsler Adult Intelligence Scale" (WAIS), a standard intelligence test for adults, and added a supplementary list from the Dutch version of the game Trivial Pursuit. For children, the questions were partly taken from a Dutch version of the "Wechsler Intelligence Scale for Children" (WISC). Again, we only selected those questions that allowed for a single word answer, and supplemented these with questions

from the Dutch version of Trivial Pursuit for children (e.g., How much is a dozen? Who discovered America?). For both participant groups, questions were selected in such a way that we could elicit different types of responses: both answers and non-answers, and both certain and uncertain responses. Both lists of questions were tested informally with 3 adults and 3 children, and indeed gave rise to the intended variety of responses. Before the actual experiment, participants were told that the questions ranged in level of difficulty, and that we did not expect them to be able to answer all questions correctly. Both adult and child speakers were always given the list of questions in one of two random orders. The appendix contains the lists of questions used in the respective experiments.

Experimental procedure Child and adult speakers underwent the same three-step procedure, modulo some small differences detailed below.

First, speakers were asked the series of questions by the experimenter, and the speakers' responses were filmed using a digital camera. The experimenter asked the series of questions one by one, and the pace of the experiment was determined by the participant. The experiment was set up in such a way that participants could not see the experimenter (who sat behind a screen in the same room), to prevent participants from picking up any cues from the experimenter which they might interpret as feedback about the (in)correctness of a given answer. Participants were told about this motivation for the screen and they were informed that the experimenter could see them via the digital camera on a computer screen that was also positioned behind the screen, thereby motivating the presence of the camera.

Second, after this test, the same sequence of questions was presented again, but now participants only had to indicate how sure they were that they would recognize the correct answer if they would have to find it in a multiple-choice test. Third, and finally, the same sequence of questions was presented yet again, but now in a multiple-choice paper-and-pencil test in which the correct answer was mixed with three plausible alternatives. For instance, in the child experiment the question "What is the name of

restaurants where you can order a happy meal?" listed McDonalds (correct) with three other fast food chains: Pizza Hut, Burger King and Kentucky Fried Chicken. For the multiple-choice recognition test, participants were instructed to answer every question, even if they had to guess. In all parts of the experiment, the questions were presented in the same random order. In none of the phases did participants receive any feedback about the (in)correctness of their answers.

The scores obtained in the second phase are referred to as the Feeling of Knowing (FOK) scores. Adult participants indicated their Feeling of Knowing on a 7-point Likert scale. For second grade children, a standard 7-point Likert scale might cause problems, hence we opted for a 5-point Likert scale using a facial representation of the items with the mouth changing from a sad face (mouth corners pulled down) to a smiling one (mouth corners pulled up). Facial representations of Likert scales are fairly standard for testing with children and are used, for instance, by Lockl and Schneider (2002) in their Feeling of Knowing studies, but have also been used for the evaluation of the usability of educational software (e.g., Read et al. 2000), to study children's perception of irony (Harris & Pexman 2003) and to measure subjective pain experience in children (e.g., Bieri et al. 1990).

Both the adult and child participants started all 3 phases with a training part to make them acquainted with the task and the stimuli (the questions), during which the experimenter was not behind the screen and interacted directly with the participants. For adults, three questions (different from the 40 test questions) were used for this purpose. For children, a longer list (10 questions, different from the 30 test questions) was used, to reduce the chances of misunderstanding to a minimum. The 10 questions for child participants started with very simple ones (e.g., the first training question was "What do we call *this* finger?", where the experimenter raised her thumb), but also included more difficult questions to illustrate the variety of question levels. During the second and third phase, the 10 training questions were used to explain the respective questionnaires. For the second phase, child participants were asked to indicate whether they thought they could recognize the correct answer (e.g., thumb) if they were given

four finger names (thumb, middle finger, ring finger, little finger). In all three phases, 10 training questions proved to be sufficient to guarantee that child participants could perform the test with the 30 questions independently and adequately.

For the purpose of comparison, both Likert scales were recoded to the interval [0,1], with 0 = "I will absolutely not recognize the correct answer in a multiple choice test" and 1 = "I will definitely recognize the correct answer in a multiple choice test". A speaker's utterance is said to be **uncertain** if the corresponding FOK score is low, and **certain** if the FOK score is high.

To stimulate participants to do their best and guess the correct answer in case of uncertainty, they were told that the "winner" of the game (the person with most answers correct in the first test) would receive a small award (a book token for the adults, and a bag of sweets for the children). In addition, all children received a small reward (a lollipop).

As an illustration, consider 4 actual responses from the child experiment (translated from Dutch) to the question "Who discovered America?":

a. Columbus;

b. Saddam Hussein;

c. Pirates;

d. I don't know.

This example shows cases of a correct answer (a), two incorrect ones, also referred to as "commission errors" (b and c) and, finally, a non-answer, or "omission error" (d).

Labeling and annotation All utterances from the first test (both by adults and by children) were transcribed orthographically and manually labeled with a number of auditory and visual features by four independent transcribers on the basis of an explicit labeling protocol. The presence or absence of the following verbal and visual features was labeled:

Filler Whether the utterance contained fillers (‘uh’, ‘uhm’), or whether these were absent.

Delay Whether a speaker responded immediately, or took some time to respond.

High intonation Whether a speaker’s utterance ended in a high boundary tone or not.

Eyebrow movement Whether one or more eyebrows departed from neutral position during the utterance or not.

Smile Whether the speaker smiled (even silently) during the response or not.

Funny face Whether the speaker produced a ‘marked facial expression’ or not.

All features were labeled categorically (in terms of presence and absence), and no distinction was made between one or more occurrences of any of the features. Labelers were always blind to the condition; all features were labeled independently from the FOK scores to avoid circularity. Features were only marked if they were clearly present, and only based on perceptual judgments.

The three auditory features were also studied by Smith and Clark (1993) and Brennan and Williams (1995). Since Brennan and Williams (1995) found little difference between different kinds of fillers, we did not differentiate between ‘uh’, ‘uhm’ or ‘mm’. We did not attempt to isolate question intonation, as it turned out to be difficult to consistently differentiate ‘real’ question intonation from list intonation. We did not measure the actual length of the delays.

The three visual features are roughly comparable with some of the Action Units (AUs) described by Ekman and Friesen (1978) for their *Facial Action Coding System*, which builds on the assumption that basic facial actions can be described in terms of single muscular actions, and more complex facial expressions can be described using these atomic building blocks. Of the three visual features under consideration here, smiling is related to AUs 12 and 13 and eyebrow raising to AUs 1 and 2. Funny faces typically consist of a combination of AUs such as lip corner depression (AU 15), lip stretching (AU 20) or lip pressing (AU 24), combined with eye widening (AU 5) and

possibly brow movement as well. See Figures 1 and 2 for representative examples of the visual cues for both adult and child speakers.

FIGURE 1 APPROXIMATELY HERE.

FIGURE 2 APPROXIMATELY HERE.

On the basis of a preliminary labeling protocol, utterances from two adult and two child speakers were labeled collectively by the four annotators. This collective effort revealed that for most features the labeling was unproblematic. For the few difficult cases a consensus labeling was reached after discussion. This first phase resulted in an explicit labeling protocol with various reference instances for the relevant features, after which the labelers proceeded individually by labeling one or two cues in the recordings from the remaining speakers, so that all utterances from all speakers were coded in terms of all features.

Statistical analysis Our statistical analysis procedure closely follows the ones proposed by Smith and Clark (1993) and Brennan and Williams (1995). We obtained 800 adult responses, 40 from each of the 20 participants, and 630 children responses, 30 from 21 participants. As already remarked by Smith and Clark, these responses are not independent so that analyses across all responses would be inappropriate. Therefore, the following tests are always based on individual analyses per speaker. We computed correlation coefficients for each participant individually, transformed the correlations for both adults and children into Fischer's z_r scores, and tested the average z_r score against zero. The average z_r scores were then transformed back into correlations for reporting in this article. Similarly, when comparing means, we computed a mean for each speaker, and used these composite scores in our analyses. In any individual analysis, we did not include any participant for whom we could not compute an individual correlation or mean, so some of our statistics, as in Smith and Clark (1993), are based on a total n of less than 20 (adults) or less than 21 (children). The ANOVA tests reported below compare both means for participants, and for items.

Results

TABLE 1 APPROXIMATELY HERE.

Table 1 shows the average FOK scores for adults and children as a function of different response categories. The first thing to note is that the overall FOK scores per category are strikingly similar, the only difference being that children assign lower FOK scores to incorrect answers than adults. Adults' mean FOK ratings were higher when they were able to produce an answer than when they were not (with participants as random factor, $F1_{(1,17)} = 71.821, p < .001$; with items as random factor, $F2_{(1,23)} = 59.028, p < .001$). The same outcome is true for the children's data (with participants as random factor, $F1_{(1,19)} = 134.617, p < .001$; with items as random factor, $F2_{(1,20)} = 126.971, p < .001$). The mean FOK ratings were higher for correctly recalled answers in the adults' data than for the incorrect ones (with participants as random factor, $F1_{(1,19)} = 149.233, p < .001$; with items as random factor, $F2_{(1,35)} = 38.086, p < .001$). Again, this was similar for the children (with participants as random factor, $F1_{(1,20)} = 111.037, p < .001$; with items as random factor, $F2_{(1,23)} = 77.101, p < .001$). Adults also gave higher FOK ratings on average for responses that they later recognized in the multiple choice test (with participants as random factor, $F1_{(1,18)} = 55.018, p < .001$; with items as random factor, $F2_{(1,27)} = 19.714, p < .001$). This again was true also for the children's data (with participants as random factor, $F1_{(1,20)} = 92.999, p < .001$; with items as random factor, $F2_{(1,16)} = 6.603, p < .05$). These data thus show that there is a close correspondence between the FOK scores and the correctness or incorrectness of a response in both the open test and the multiple-choice. The results of both adults and children are similar to those of Smith and Clark (1993) and Brennan and Williams (1995). Note that these findings support the so-called 'trace-based' view of memory access (see e.g., Koriat 1993, Lockl & Schneider 2002). According to this view the FOK scores for incorrect answers (commission errors) should be relatively high (but lower than those for correct answers), and they should also be higher than those for non-answers (omission errors), which is the case both for adults and children.

TABLE 2 APPROXIMATELY HERE.

TABLE 3 APPROXIMATELY HERE.

Since no systematic differences between male and female participants were found in the analyses, the results are collapsed across gender. Tables 2 and 3 display the labeling results for adult answers and non-answers respectively, by comparing average FOK scores for utterances in which a specific marked feature is present with those for utterances in which it is absent. Table 2 shows that the presence of a marked verbal or visual feature in answers coincides with a lower FOK score (significantly for filler, delay, high intonation, eyebrow and funny face, but not for smile). Table 3, on the contrary, shows that the presence of a marked feature in non-answers generally leads to higher FOK scores, although the differences between presence and absence of a feature are only significant for filler and delay, probably because of the relatively limited number of data points here. Notice also that non-answers ("I don't know") are arguably inherently less likely to be uttered with a high intonation, presumably because speakers do not need to question their own internal state (see e.g., Geluykens 1987), which is reflected in a zero difference score.

Tables 4 and 5 describe the labeling results for child answers and non-answers. The picture for child answers is similar to that for the adults, but the results are overall less pronounced. Looking at the scores for answers in Table 4, it can be seen that in most cases the presence of a verbal or visual feature is associated with a lower FOK score, albeit that the differences are generally small (except for delay and funny face) and not significant for filler and smile. It is particularly surprising that fillers (a strong cue for adult uncertainty) play only a marginal role for uncertainty signaling in children. Table 5 shows the results for the child speakers' non-answers, and does not offer any significant differences. The general picture is not clear either, in contrast to the adult results for non-answers, since the presence of some features (filler, delay and smile) leads to marginally higher FOK scores, while the presence of other features (eyebrow and funny face) leads to marginally lower FOK scores.

TABLE 4 APPROXIMATELY HERE.

TABLE 5 APPROXIMATELY HERE.

In order to learn more about the cue value of *combinations* of features, we also calculated, for answers and non-answers separately, the average FOK scores for responses that differ regarding the number of marked feature settings (the sum of the six audiovisual features). These correlations are given in Table 6, which again illustrates opposite trends for the two response categories: for answers, the average FOK score decreases, in both the adult and child data, with an increasing number of marked features, while the opposite is true for non-answers, though this effect is very small and non-significant in the child data.

TABLE 6 APPROXIMATELY HERE.

Discussion

In the first experiment, Hart's (1965) FOK paradigm was used to elicit certain and uncertain utterances from adult and child speakers. From the labeling analysis, it appears that particular audiovisual surface forms of the utterances produced by the adult speakers are indicative of the amount of confidence they have about the correctness of their response. For answers, lower FOK scores correlate significantly with the presence of delays, filled pauses, high intonation, eyebrows, and funny faces. For non-answers, the relationships between FOK scores and the different audiovisual features is the mirror image of the outcome with answers, but the differences between means are only significant for filler and delay, probably due in part to the limited number of data points. Additionally, as argued above, it seems plausible that non-answers are inherently less likely to be uttered with a high intonation. Arguably, a speaker who utters a low FOK non-answer is relatively certain that he or she does not know the answer, and hence does not need to continue a longer memory search in an attempt to retrieve an answer. These results generalize the earlier finding of Smith and Clark (1993) that answers and non-answers differ in speaker behavior. Interestingly, the overall picture for child

speakers is somewhat similar but much less pronounced than that for adults. For child answers, lower FOK scores generally correlate with the presence of audiovisual cues, with the exception of smile and (strikingly) fillers. For child non-answers, the relation between FOK scores and audiovisual features reveals no clear picture and no significant results. In sum, it seems fair to conclude that our adult speakers use audiovisual cues more often and more consistently to signal their level of certainty than our child speakers.

The first study focussed on speakers' production of uncertainty, to gain insight into audiovisual correlates of Feeling of Knowing. In the second study we investigate the perceptual relevance of such features. For this, we use earlier work by Brennan and Williams (1995) on Feeling of Another's Knowing as our main source of inspiration. The novelty lies in the fact that we obtain perceptual data from both child and adult judges looking at both child and adult speakers.

EXPERIMENT II: DETECTING UNCERTAINTY

Method

Participants 80 native speakers of Dutch participated as judges, 40 adults and 40 children (20 male and 20 female per group), and all different from the speakers that participated in the production studies. The children were in two different second grade classes from the St. Jozef School, all children were 7 or 8 years old. The adults were all students from Tilburg University, between 18 and 25 years.

Stimuli From the corpus of answers collected in the first study, we selected 30 adult utterances and 30 child utterances, both with an equal distribution of high FOK and low FOK scores. The selection was based on the written transcriptions of the responses. Given the individual differences in the use of the FOK scale, we chose to use —per speaker— the highest score as an instantiation of high FOK and the two lowest as representations of low FOK. The original selection of stimuli was random, but utterances

were iteratively replaced until the following criteria were met:

1. the original question posed by the experimenter should not re-appear in the speakers' response;
2. all the answers should be lexically distinct; and
3. there should be maximally two answers per speaker.

Having applied this procedure on the basis of written transcriptions of the data, we finally replaced some stimuli, if the signal-to-noise ratio made them unsuitable for the perception experiment. Initially we planned to include non-answers in the perception experiment as well, but since we could not find sufficiently many high FOK non-answers meeting the criteria among the children's responses we had to give up this intention. Table 7 summarizes the selection of the stimuli.

TABLE 7 APPROXIMATELY HERE.

Experimental design The experiment had a 2 x 2 balanced Latin square design with original FOK score as a within participants factor and Speaker and Judge as between participants factors. Thus, of the 40 adult judges, 20 judged stimuli from adult speakers, and 20 from child speakers, and likewise for the 40 child judges. See Table 8 for a schematic representation of the design.

TABLE 8 APPROXIMATELY HERE.

Experimental procedure Stimuli were presented on a screen where the judges first saw the stimulus ID (1 through 30) and then the actual stimulus. The inter-stimulus interval was 3 seconds for adult judges and 6 seconds for child judges. For all four groups, the experiment was preceded by a short exercise session to acquaint judges with the kinds of stimulus materials and the procedure. Judges were instructed to estimate whether

speakers were uncertain about their answers or not. Adult judges scored this on a 7-point Likert scale, child judges on a 5-point Likert scale (using the same facial representation as above). For presentation purposes, both scales are recoded to the interval [0, 1], with 0 = "very uncertain" and 1 = "very certain". These scores are referred to as the Feeling Of Another's Knowing (FOAK) scores (Jameson et al. 1993, Brennan & Williams 1995).

Results

TABLE 9 APPROXIMATELY HERE.

The overall results are summarized in Table 9. All tests for significance were done using analysis of variance (ANOVA). There was a small but significant main effect of speaker ($F_{(1,76)} = 6.574, p < .05$). This means that overall child speakers received slightly higher FOAK scores than adult speakers (0.59 and 0.56 respectively). Additionally, a main effect of FOK score was found ($F_{(1,76)} = 601.987, p < .001$). As one would expect, stimuli with a high FOK score get overall higher FOAK scores than stimuli with a low FOK.

To see whether there are differences between adults and children, we have to look at the interactions. A significant two-way interaction was found between speaker and FOK score ($F_{(1,76)} = 26.281, p < .01$). Inspection of Table 9 reveals that stimuli from adult speakers receive more diverging FOAK scores than stimuli from child speakers, irrespective of who the judges are. In other words, overall, the FOAK scores are more accurate for adult speakers than for child speakers. There was also a significant interaction between judge and FOK score ($F_{(1,76)} = 62.726, p < .01$); differences between FOAK scores for high and low FOK stimuli are larger for adult than for child judges (cf. Table 10). Table 10 can also be used to interpret the significant 3-way interaction between judge, speaker and FOK ($F_{(1,76)} = 19.419, p < .01$), since it shows that the FOAK scores for high FOK and low FOK stimuli differ most for adults judging adults and least for children judging children.

TABLE 10 APPROXIMATELY HERE.

Discussion

The experiment revealed a number of clear differences between adults and children, both as speakers and as judges. Overall, we found that FOAK scores assigned to stimuli from adult speakers offer a more accurate reflection of the original FOK score than stimuli from child speakers. This is not unexpected: what is not clearly signalled can not be detected, and the results from the first experiment already revealed that adult speakers are more systematic in cuing their (un)certainty. The results of the current experiment also indicate, somewhat surprisingly, that adults are "better" judges of uncertainty than children. It is unclear why this is the case: it might be that the children in our group still have a somewhat underdeveloped theory of mind related to uncertainty, or alternatively, it might be that their understanding of uncertainty signalling is not fully developed yet. It is worth stressing that the studies mentioned in the introduction, showed that children of age 4 are generally capable of distinguishing certain and uncertain utterances, provided the certainty level was *explicitly* marked, whereas in the current experiment uncertainty was marked implicitly (non-verbally).

GENERAL DISCUSSION

We have described two experiments on signaling and detecting uncertainty in audiovisual speech by adults (20 - 50 years old) and second grade school children (7-8 years old).

In the first study (production of uncertainty), both adult and child participants were asked a series of factual questions and their answers were filmed. It was interesting to observe that adults and children gave highly similar Feeling of Knowing scores, with highest FOK scores for correct answers (no errors), somewhat lower FOK scores for incorrect answers (commission errors) and lowest scores for non-answers (omission errors). This suggests that the meta-memory assessment for this particular task operates

in a similar way for both groups of participants, which was as expected and is consistent with earlier work by for instance Lockl and Schneider (2002). The analysis of the recordings (in terms of uncertainty cues proposed by Smith & Clark 1993, Brennan & Williams 1995, Swerts & Krahmer 2005) revealed some interesting differences in the audio signalling of uncertainty. It was found that for adult speakers the occurrence of prosodic features such as fillers, delays and high intonation is a clear cue for low FOK answers and for high FOK non-answers. Even though a somewhat similar trend can be observed in the child data, this only yields significant differences for delay and high intonation, and only for the answers. It is interesting to see that fillers appear to have no relation with uncertainty in the child data, while it is one of the clearest cues for adults. It would be very interesting to perform a separate, more controlled perception study along the lines of Brennan and Williams (1995) to check whether participants in the child age group are sensitive to fillers for their FOAK judgments.

For the visual cues the situation is somewhat more complex: for adult speakers the presence of a marked visual feature (eyebrow movement, smiling or making a 'funny face') systematically corresponds with lower FOK scores in the case of answers (for the non-answers no significant differences were found, although the trends in all cases were in the expected, reverse direction). For children, only eyebrow and funny face play a role, in that the presence of either of these cues corresponds with a lower FOK score. Interestingly, this holds both for answers (in line with the adult results) and (albeit non-significantly) non-answers, in contrast with the adult results. In sum, while no meta-cognitive differences were found between adults and children, it appears that our adult speakers use audiovisual prosody to signal their level of uncertainty in a more consistent and more clear way than our child speakers.

In the second study (perception of uncertainty), adults and children judged the uncertainty (or the Feeling of Another's Knowing) when looking at adult and child speakers. From this study, we may conclude that both child and adult judges give more accurate FOAK scores (i.e., make better estimates of a speaker's level of certainty) for answers from adult than from child speakers. This is in line with the findings of

the first study ("what is not signalled cannot be detected"). The second study also revealed that, contrary to our expectations, child judges overall provide less accurate FOAK scores than adult judges, possibly because their understanding of uncertainty and/or the signalling thereof is still underdeveloped. More research is needed to find an adequate explanation for this. In general, the results of the perception experiments suggest that our second grade children make less effective use of the audiovisual cues than adults, which is in line with some earlier findings (Cruttenden 1974, 1985; Vogel & Raimy 2002; Ackerman 1983) that school age children make less effective use of subtle auditory cues than adults do.

FIGURE 3 APPROXIMATELY HERE.

These outcomes raise an interesting question: why is it that children between 7 and 8 years old, who seem to be capable of forming meta-cognitive judgments about their own certainty and uncertainty, do not signal this uncertainty in the way that adults do? It might be that the child participants were more easily distracted or that they are less concerned with self-presentation than the adult participants. Even though both factors may play a role, we conjecture on the basis of inspection of the recordings that self-presentation is the more important one (if the children were distracted, one would expect them, for instance, to 'forget' the question, which hardly ever happened). With regard to self-presentation adult speakers clearly indicate while answering, whether they are certain about the correctness of their answer or not. In doing so, they can save face when an answer turns out to be incorrect; it is a signal to the questioner saying "this is how certain I think you can be about my answer." As mentioned, children in the age group 7-8 appear to be less worried about self-presentation in answering questions. As one piece of anecdotal evidence supporting this view we observe that some of the child participants displayed behaviors that detract from self-presentation, while none of our adult speakers exhibited these behaviors. Specifically, one child repeatedly picked her nose during the experiment, while at least two others pulled down their mouth corners with the fingers (cf. Figure 3).

In a similar vein, various child participants occasionally spent *very* long times on memory search (more than 30 seconds), without accounting for this delay in any way (and usually still providing an answer that was relevant to the question). The adult participants, by contrast, even though their search space is obviously much larger, hardly ever took more than 10 seconds to answer and in the case of prolonged search almost always indicated this to the questioner via one or more audiovisual cues. In this context it is interesting to observe that delay is a significant cue for child uncertainty while filled pauses are not. A child who does not know an answer immediately tends to engage in a longer memory search, but does not signal this to the listener as an adult would. We are not claiming that our children are slower in their search or that they really need this longer search, but rather that our adult participants would sooner abandon the search (which might suggest better meta-cognitive estimations in this respect) and that the adults signal this.

A few years ago there was a popular Dutch radio quiz ("Zeg eens uh …", *Say uh* …) which explicitly addressed adult filler usage. In this telephone quiz, participants were asked a series of factual questions, mostly related to their personal life, and were *not* allowed to use fillers. The winner was the person that managed to avoid "uh" and "uhm" for the longest non-interrupted stretch of time. Interestingly, most participants had difficulty with prolonged filler avoidance, and typically did not notice themselves when they failed do so. This suggests that filler usage in adults is in part an automatic process. The results of the Feeling of Knowing study suggest that filler usage is something that is gradually learned, and not yet fully developed in 2nd graders (which is consistent with the suggestions from MacWhinney & Osser 1977, Narayanan & Ptamiano 2002, Esposito et al. 2004, Montanari et al. 2004). Why this is the case, and when children do develop adult filler strategies are interesting questions for future research.

ACKNOWLEDGEMENTS

The research described in this paper was conducted as part of the VIDI-project “Functions Of Audiovisual Prosody (FOAP)”, sponsored by the Netherlands Organisation for Scientific Research (NWO), see `foap.uvt.nl`, and as part of the IMIX-project “Interactive Multimodal Output Generation (Imogen)”, also sponsored by the Netherlands Organisation for Scientific Research (NWO). Swerts is also affiliated with the Flemish Fund for Scientific Research (FWO-Flanders) and Antwerp University. Many thanks to Judith Schrier (Antwerp) and Jorien Scholze, Kim Smulders and Nicole Hobbelen (Tilburg) for their help in carrying out the experiments, and to Lennard van de Laar and Pashiera Barkhuysen for their help with the annotation and the experimental set-up.Thanks to Carel van Wijk and to Annemarie Krahmer-Borduin for statistical and methodological assistance. We greatly benefited from comments from Dominic W. Massaro and Laurel Fais on a previous version of this paper.

Appendix

English translations of the Dutch **child** questions used in the first experiment, as they were presented to participants in one of the random orders:

1. What do we call the young of a cow?
2. Which friend of Asterix can eat an entire wild bear?
3. What do we call a story that starts with “Once upon a time...”?
4. From which material is leather made?
5. What do we a call a building where you can borrow books?
6. What is the biggest mammal in the world?
7. George Bush is the president of which country?
8. To whom do we address the song “Kom maar binnen met je knecht?”

9. What do we call the long kinds of bread people eat in France?
10. What do we call two or more beds stacked on top of each other?
11. What do we call a small mammal with prickles on its back?
12. What kind of animals live in an aquarium?
13. How many people live in the Netherlands?
14. What do we call a person who checks your ticket in a train?
15. What do we call the long rest period many animals have during winter?
16. What do we call newborn lions?
17. What do we call a drawing on a person's skin?
18. What is the capital of the Netherlands?
19. Disneyland Paris lies in which country?
20. What does a tree have below the surface to extract water?
21. K3 comes from which country?
22. From which material is glass made?
23. What is the name of the restaurants where you can order a Happy Meal?
24. Which month follows March?
25. What do we call a car which brings ill people to the hospital?
26. What do we call the water drops that fall down out of the clouds?
27. How many days are there in a week?
28. How much is a dozen?
29. Who discovered America?

30. What do we call a bird that can talk?

English translations of the Dutch **adult** questions used in the first experiment, as they were presented to participants in one of the random orders:

1. What does one call the sticks used in golf?
2. Who made the drawings for “Jip and Janneke”?
3. The sahara lies in which continent?
4. Which novel about a knight is the most reprinted book after the Bible?
5. How many months does it take the moon to circle the earth?
6. What does the abbreviation ‘Fl’ for the Dutch guilder stand for?
7. What is the largest mammal?
8. What is the name of the gang of robbers that terrorized Limburg in the 18th century?
9. Who, according to legend, was the bishop of Myra?
10. In which Dutch quiz show are the contestants awarded with a toy monkey for each good answer?
11. What is the highest mountain of the Alps?
12. Who wrote Faust?
13. What is the chemical symbol of water?
14. What does the word ‘Jihad’ mean?
15. What color of light is used on the starboard side of a boat?
16. What is Rembrandt’s last name?
17. Which television series is about the Forrester and Spectra families?

18. Guide Gezelle was a famous man. What was his occupation?
19. What is the boiling temperature for water?
20. In which wind-direction does one travel from Amsterdam to Brussel?
21. What is the name of the cartoon character who owns Pluto?
22. Egypt lies in which continent?
23. Who is the head of state of the Vatican?
24. Who wrote “The discovery of heaven”?
25. What is a “Friese doorloper”?
26. Brazil lies in which continent?
27. What is the pseudonym of the Mexican Don Diego de la Vega?
28. Who wrote Hamlet?
29. Which rocker is also known as “The King”?
30. How many darts is a player allowed to throw in one turn?
31. In which wind-direction does one travel from London to Berlin?
32. Which disease was known during the Middle Ages as “The Black Death”?
33. What is the capital of Switzerland?
34. Supporters of which football club sing “Geen woorden maar daden”?
35. How many degrees are in a circle?
36. Approximately, how many people live in the Netherlands and Belgium?
37. In which country did the Inca’s live?
38. Which person from the Bible went to look for mustard?

39. Which Dutch soap series has been running on television for the longest period?

40. Who wrote the Iliad?

References

Ackerman, B. (1983), Form and function in childrens' understanding of ironic utterances, *Journal of Experimental Child Psychology*, **35**, 487–508.

Barkhuysen, P., Krahmer, E., & Swerts, M. (2005a), Predicting end of utterance in mulitmodal and unimodal conditions, submitted.

Barkhuysen, P., Krahmer, E., & Swerts, M. (2005b), Problem Detection in Human-Machine Interactions based on Facial Expressions of Users, *Speech Communication*, **45**(3), 343-359.

Bartsch, K., & Wellman, H (1995), *Children talk about the mind*, New York: Oxford University Press.

Beach, C., Katz, W., & Skowrosnki, A. (1996), Children's processing of prosodic cues for phrasal interpretation, *Journal of the Acoustical Society of America*, **99**, 1148–1160.

Bieri, D., Reeve, R., Champion, G., Addicoat, L. & Ziegler, J. (1990), The Faces Pain Scale for the self-assessment of the severity of pain experienced by children: development, initial validation, preliminary investigation for ratio scale properties, *Pain*, **41**, 139–150.

Bisanz, G., Vesonder, G., & Voss, J. (1978), Knowledge of one's own responding and relation of such knowledge to learning: A developmental study, *Journal of Experimental Child Psychology*, **25**, 116–128.

Brennan, S.E., & Williams, M. (1995), The feeling of another's knowing: prosody and filled pauses as cues to listeners about the metacognitive states of speakers, *Journal of Memory and Language*, **34**, 383-398.

Brown, A., & S. Lawton (1977), The feeling of knowing experience in educable retarded children, *Developmental Psychology*, **13**, 364–370.

Butterfield, E., Nelson, T., & Peck, V. (1988), Developmental aspects of the feeling of knowing, *Developmental Psychology*, **24**, 654–663.

Cassell, J. (2004), Towards a Model of Technology and Literacy Development: Story Listening Systems, *Journal of Applied Developmental Psychology*, **25** (1), 75–105.

Clark, H., & Krych, M. (2004), Speaking while monitoring addressees for understanding, *Journal of Memory and Language*, **50**, 62–81.

Cruttenden, A. (1974), An experiment involving comprehension of intonation in children from 7 to 10, *Journal of Child Language*, **1**, 221-231.

Cruttenden, A. (1985), Intonation comprehension in ten-year-olds, *Journal of Child Language*, **12**, 643–661.

Cruttenden, A. (1986), *Intonation*, Cambridge: Cambridge University press.

Cultice, J., Somerville, S., & Wellman, H. (1983), Preschoolers' memory monitoring: Feeling-of-knowing judgments, *Child development*, **54**, 1480–1486.

Doherty-Sneddon, G., & Kent, G. (1996), Visual signals and the communication abilities of children, *Journal of Child Psychology and Psychiatry*, **37**, 949–959.

Ekman, P. (1979), About brows: Emotional and conversational signals, in: *Human Ethology*, M. von Cranach et al. (eds.), Cambridge University Press, Cambridge, pp. 169–202.

Ekman, P., & Friesen, W.V. (1978). *The Facial Action Coding Scheme*, Palo Alto: Consulting Psychologists' Press.

Esposito, A., Marinaro, M., & Palombo, G. (2004), Children speech pauses as markers of different discourse structures and utterance information content, in: *From Sound to Sense*, June 11-13, MIT, Cambridge, MA, pp. 139–144.

Flavell, J. (1999), Cognitive development: Childrens' knowledge about the mind, *Annual Review of Psychology*, **50**, 21-40.

Geluykens, R. (1987), Intonation and speech act type. An experimental approach to rising intonation in declaratives. *Journal of Pragmatics*, **11**, 483–494.

Graham, T. (1999), The role of gesture in children's learning to count, *Journal of Experimental Child Psychology*, **74**, 333-355.

Granström, B., House, D., & Lundeberg, M. (1999), Prosodic cues in multimodal speech perception, *Proceedings 14th International Conference of the Phonetic Sciences (ICPhS)*, San Francisco.

Harris, M. & P. Pexman (2003), Children's perceptions of the social function of verbal irony, *Discourse Processes*, **36**, 147–165.

Hart, J.T. (1965), Memory and the feeling-of-knowing experience, *Journal of Educational Psychology*, **56**, 208–216.

Jameson, A., Nelson, T., Leonesio, R., & Narens, L. (1993), The feeling of another person's knowing, *Journal of Memory and Language*, **32**, 320 – 335.

Keating, P., Baroni, M., Mattys, S., Scarborough, R., Alwan, A., Auer, E., & Berstein, L. (2003). Optical phonetics and visual perception of lexical and phrasal stress in English, in: *Proceedings 16th International Conference of the Phonetic Sciences (ICPhS)*, Barcelona, Spain, pp. 2071–2074.

Keenan, T., & Quigely, K. (1999), Do young children use echoic information in their comprehension of sarcastic speech? *British Journal of Developmental Psychology*, **17**, 83–96.

Koenig, M., Clément, F., & Harris, P. (2004), Trust in testimony; Children's use of true and false statements, *Psychological Science*, **15**(10), 694–698.

Koriat, A. (1993). How do we know that we know? The accessibility model of the feeling of knowing, *Psychological Review*, **100**, 609–639.

Krahmer, E., & Swerts, M. (2004), More about brows, in: *From Brows to Trust: Evaluating Embodied Conversational Agents*, Zs. Ruttkay & C. Pelachaud (eds.), Kluwer Academic Press, Dordrecht.

Kreutzer, M., Leonard, S., & Flavell, J. (1975), An Interview Study of Children's Knowledge about Memory, *Monographs of the Society for Research in Child Development*, **40**, 1–57.

Ladd, D.R. (1996), *Intonational Phonology*, Cambridge: Cambridge University Press.

Leuckefeld, K., Hahne, A., & Alter, K. (2003), Neuronal correlates of prosodic information processing and syntactic disambiguation in school-aged children, *Paper presented at CUNY Conference on Human Sentence Processing*, MIT.

Lockl, K., & Schneider, W. (2002), Developmental trends in children's feeling-of-knowing judgments. *International Journal of Behavioral Development*, **26**, 327–333

Matsui, T., McCagg, P., Yamamoto, T., Murakami, Y., & Fernald, A. (2003), Japanese Preschoolers' Early Understanding of (Un)certainty: A Cultural Perspective on the Role of Language in Development of Theory of Mind, in *Proceedings of the 28th annual Boston University Conference on Language Development (BU-CLD)*, Brugos, A., Micciulla, L., and Smith, C. (eds.), Boston: Cascadilla Press, pp. 350–362

MacWhinney, B., & Osser, H. (1977), Verbal planning functions in children's speech, *Child Development*, **48**, 978–985.

McNeill, D. (1992), *Hand and Mind: What gestures reveal about thought*, Chicago: The University of Chicago Press.

Montanari, S., Yildirim, S., Khurana, S., M. Landes, Lawyer, L., Andersen, E., & Narayanan, S. (2004), Analyzing the interplay between spoken language and gestural cues in conversational child-machine interactions in pre/early literare

age groups, in: *Proceedings of InSTIL/ICALL - NLP and Speech Technologies in Advanced Language Learning Systems*, Venice, 17-19 June.

Moore, C., Bryant, D., & Furrow, D. (1989), Mental terms and the development of certainty. *Child Development*, **60**, 167–171.

Moore, C., Pure, K., & Furrow, D. (1990), Children's understanding of the modal expression of speaker certainty and uncertainty and its relation to the development of a representational theory of mind, *Child Development* **61**, 722–730.

Munhall, K., Jones, J., Callan, D., Kuratate, T., & Vatikiotis-Bateson, E. (2004), Visual prosody and speech intelligibility, *Psychological Science*, **15**, 133–137.

Narayanan, S., & Potamianos, A. (2002), Creating conversational interfaces for children, *IEEE Transactions on Speech and Audio Processing*, **10**, 65–78.

Nelson, T. (1993), *Metacognition: Core Readings*, Prentice Hall.

Nelson, T., & Narens, L. (1990). Metamemory: A Theoretical Framework and New Findings, *The Psychology of Learning and Motivation*, **26**, 125–141.

Pena-Brooks, A., & Hegde, M. (2000), *Assessment and Treatment of Articulation and Phonological Disorders in Children*, Texas: pro-ed.

Pressley, M. Levin, J.R., & Ghatala, E.S. (1984). Memory strategy monitoring in adults and children, *Journal of Verbal Learning and Verbal Behavior*, **23**, 270–288.

Read, J., MacFarlane, S., & Casey, C. (2002), Endurability, engagement and expectations, in: *Interaction Design and Children*, M. Bekker, et al. (eds.), Shaker Publishing.

Robinson, E., Mitchell, P., & Nye, R. (1995), Young children's treating of utterances as unreliable sources of knowledge, *Journal of Child Language*, **22**, 663–685.

Sabbagh, M., & Baldwin, D., (2001), Learning words from Knowledgeable versus Ignorant Speakers: Links between preschoolers' Theory of Mind and Semantic Development, *Child Development*, **72**(4), 1054–1070.

Schmidt, K., & Cohn, J. (2002), Human facial expressions as adaptations: Evolutionary questions in facial expression, *Yearbook of Physical Anthropology*, **44**, 3–24.

Schneider, W. (2001), Memory development in children, in N. Smelser & P.B. Baltes (eds.), *International Encyclopedia of the Social and Behavioral Sciences*, New York: Pergamon.

Schneider, W., & Lockl, K. (2002), The development of metacognition and knowledge in children and adolescents, in T. Perfect & B. Schwartz (Eds.), *Applied Metacognition* Cambridge: Cambridge University Press, pp. 224-257.

Smith, V.L., & Clark, H.H. (1993), On the course of answering questions, *Journal of Memory and Language* **32**, 25–38.

Srinivasan, R., & Massaro, D. (2003), Perceiving prosody from the face and voice: Distinguishing statements from echoic questions in English, *Language and Speech*, **46**, 1–22.

Swerts, M., & Krahmer E. (2005), Audiovisual prosody and feeling of knowing, *Journal of Memory and Language*, in press.

Vogel, I., & Raimy, E. (2002), The acquisition of compound vd. phrasal stress; the role of prosodic constituents, *Journal of Child Language*, **29**, 225–250.

Wellman, H. (1977), Tip of the tongue and feeling of knowing experiences: A developmental study of memory monitoring, *Child Development*, **48**, 13–21.

Wellman, H., Cross, D., & Watson, J. (2001), Meta-analysis of theory-of-mind development: The truth about false belief, *Child Development*, **72**(3), 655–684.

Table 1: *Average adult (N = 800) and child (N = 630) FOK scores for different response categories (there are 3 missing values in the open question part of the child experiment).*

		Adult		Child	
Experiment	Response	n	FOK	n	FOK
Open Question	All answers	704	0.90	496	0.90
	Correct Answers	575	0.94	371	0.96
	Incorrect Answers	129	0.70	125	0.54
	All non-answers	96	0.43	131	0.36
Multiple Choice	Correct Answers	717	0.88	488	0.86
	Incorrect Answers	83	0.55	142	0.56

Table 2: *Average adult FOK scores for answers as a function of presence or absence of audiovisual features. The n indicates for how many participants individual means could be computed.*

	Present (1)	Absent (2)	Diff. (1)-(2)
Filler ($n = 19$)	0.81	0.93	–0.12***
Delay ($n = 20$)	0.75	0.93	–0.18***
High Intonation ($n = 19$)	0.84	0.91	–0.07*
Eyebrow ($n = 19$)	0.83	0.92	–0.09***
Smile ($n = 17$)	0.81	0.91	–0.10
Funny Face ($n = 10$)	0.68	0.91	–0.23**

*$p < .05$; **$p < .01$; ***$p < .001$

Table 3: *Average adult FOK scores for non-answers as a function of presence or absence of audiovisual features. The n indicates for how many participants individual means could be computed.*

	Present (1)	Absent (2)	Diff. (1)-(2)
Filler ($n = 9$)	0.71	0.37	0.34**
Delay ($n = 12$)	0.60	0.31	0.29*
High Intonation ($n = 5$)	0.52	0.52	0.00
Eyebrow ($n = 8$)	0.54	0.34	0.20
Smile ($n = 12$)	0.53	0.39	0.14
Funny Face ($n = 4$)	0.54	0.41	0.13

*$p < .05$; **$p < .01$; ***$p < .001$

Table 4: *Average children FOK scores for answers as a function of presence or absence of audiovisual features. The n indicates for how many participants individual means could be computed.*

	Present (1)	Absent (2)	Diff. (1)-(2)
Filler ($n = 15$)	0.85	0.91	–0.06
Delay ($n = 20$)	0.74	0.94	–0.20***
High Intonation ($n = 20$)	0.87	0.93	–0.06**
Eyebrow ($n = 15$)	0.82	0.94	–0.12**
Smile ($n = 11$)	0.89	0.89	0.00
Funny Face ($n = 13$)	0.73	0.92	–0.26*

*$p < .05$; **$p < .01$; ***$p < .001$

Table 5: *Average children FOK scores for non-answers as a function of presence or absence of audiovisual features. The n indicates for how many participants individual means could be computed.*

	Present (1)	Absent (2)	Diff. (1)-(2)
Filler ($n = 10$)	0.44	0.42	0.02
Delay ($n = 15$)	0.40	0.36	0.04
High Intonation ($n = 11$)	0.40	0.40	0.00
Eyebrow ($n = 8$)	0.33	0.43	–0.10
Smile ($n = 12$)	0.42	0.35	0.07
Funny Face ($n = 12$)	0.42	0.45	–0.03

Table 6: *Pearson correlation coefficients of FOK scores with number of marked features for adult and child data.*

Correlations of	Adult		Children	
FOK scores with	Answers	Non-answers	Answers	Non-answers
	($n = 20$)	($n = 13$)	($n = 21$)	($n = 13$)
Marked features	–.410***	.690***	–.390***	.080

*** $p < .001$

Table 7: *Selection of stimuli.*

	FOK		
Speaker	Low	High	*Total*
Adult	15	15	30
Child	15	15	30
Total	30	30	60

Table 8: *Selection of participants.*

	Speaker		
Judge	Adult	Child	*Total*
Adult	20	20	40
Child	20	20	40
Total	40	40	80

Table 9: *Average FOAK scores for adult and child judges as a function of speaker type and FOK score.*

		FOAK scores of	
Speaker	FOK	Adult judges	Child judges
Adult	High	0.79	0.66
	Low	0.32	0.47
Child	High	0.70	0.70
	Low	0.44	0.53
Average		0.56	0.59

Table 10: *Average FOAK difference scores for adult and child judges as a function of speaker type. Differences are computed via FOAK for high FOK stimuli minus FOAK for low FOK stimuli.*

	FOAK difference for	
Speaker	Adult judges	Child judges
Adult	0.47	0.26
Child	0.19	0.17

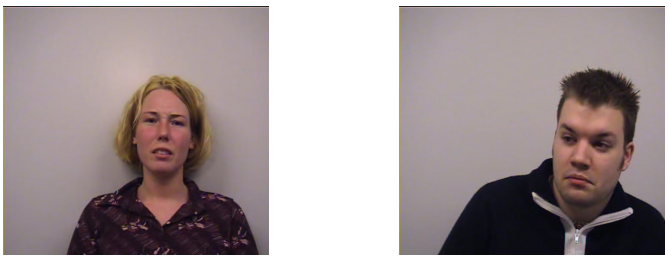
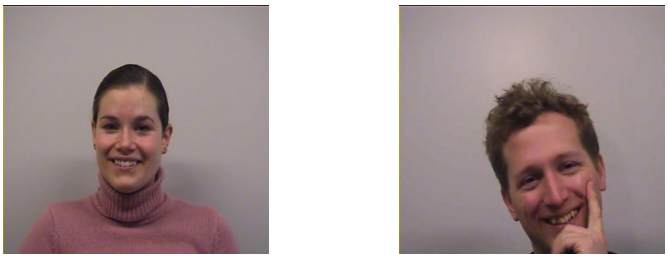
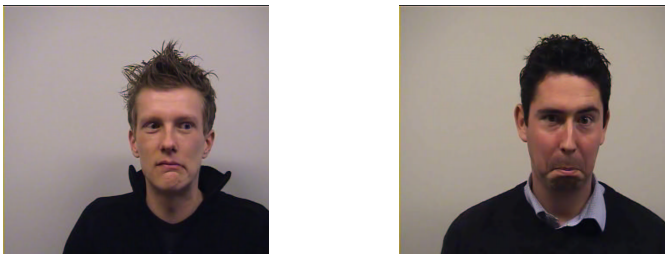
Label	Example
Eyebrow	
Smile	
Funny face	

Figure 1: *Stills from the adult experiment illustrating the three annotated visual features.*

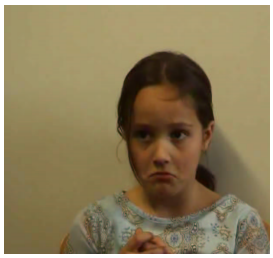
Label	Example	
Eyebrow		
Smile		
Funny face		

Figure 2: *Stills from the child experiment illustrating the three annotated visual features.*

Figure 3: Example of lack of self-presentation from the child experiment.