Chapter 2

THE VOICE TEST: A NEW ADVANCED THEORY OF MIND TASK FOR ITALIAN SCHOOL AGE CHILDREN

As I reviewed in the first chapter, researchers traditionally have been considered Theory of Mind (ToM) as a false belief understanding (Wimmer & Perner, 1983; Perner & Wimmer, 1985), assessed usually in preschool children.

In the present work I adopt a wide definition of ToM (Bruner & Feldman, 1993; Hughes & Leekam, 2004), implying a multi-componential view of it (Wellman & Liu, 2004), beyond the false-belief comprehension, that comprises a wide range of inners states, not only cognitions, but also emotions and motivational states (Astington, 2001). This wide point of view first of all allows ToM researchers to investigate ToM in infants and in older children, before and after the false belief understanding develops. I don't want underestimate nor overestimate the importance of the false belief comprehension, that certainly is an unequivocal marker of mentalistic understanding (Wellman, Cross & Watson, 2001), but I consider it just as an aspect, among others, implied in ToM.

Secondly, the wide definition adopted here brings researchers to find ecological instrument, nearer to real life than classical false belief tasks, in order to assess on the one hand ToM precursors (see paragraph 1.3.1.) in infants and on the other hand the growing mentalizing ability, that become more and more complex in school age

children. This means to use in ToM assessment the same communication "channels" we use in everyday life when we recognize mental states, using them to make sense and predict behaviors, that are the visual (i.e., what we see, such as actions and facial expressions) and auditory (i.e., what we heard) channels, besides the "cognitive" channel (it includes, for example, the memory ability, the comprehension and production of language, the executive function; e.g., we use this channel when we interpret some previous facts as cause of the present behavior and use them to understand the mental state below or when we give meaning to the words that a person says in order to understand his mental states).

For example, when we see a man frowns his brows, gesticulates animatedly (visual channel) and shouts (auditory channel) we can understand that he is probably angry. Or when we heard a woman says to a person, with a faltering voice (auditory channel), giving quick looks around (visual channel), in the middle of a busy crossing: "But I must turn right or left?" (cognitive channel), we can give meaning to her behaviors thinking that she is puzzled. Social functioning, therefore, requires the rapid processing of various stimuli conveyed principally by facial expression, vocal prosody and expression in the eyes.

Usually we use many information regarding behaviors (such as actions, verbal and non verbal behavior), that we grasp with all our senses (mainly sight and hearing), to impute mental states to the self and to the others. It is not simple to operazionalize what we experience daily when we mentalize; a solution could be to consider one aspect at a time. Researchers were focused particularly on aspects referring to what I called visual and cognitive channel. For example, some researches studied if children are able to predict, from the knowledge of the antecedent (e.g., a boy receives an unwelcome Christmas present from his parents), which mental state a story character has (e.g., he

will be sad or angry, or he will mask his real emotion and will say a lie to his parents because he doesn't want to offend them) (e.g., Happè, 1994; Baron-Cohen et al., 1999; Muris et al., 1999; Pons & Harris, 2000). Some other researches analyzed if children are able to understand mental states from visual cues, such as persons' facial expressions or schematized faces (e.g., Baron-Cohen, Wheelwright & Jolliffe, 1997; Rutherford, Baron-Cohen & Wheelwright, 2002; Golan, Baron-Cohen & Hill, 2006; Golan, Baron-Cohen, Hill & Golan, 2006).

Whereas in the field of study of ToM there are few researches focalised on the mental state understanding through the auditory channel. This is quite surprising considering that for example from early months of life infants are sensitive to vocal cues (Mastropieri & Turkewitz, 1999), respond appropriately to communication and they are able to grasp and differentiate emotions only hearing vocal cues (Caron, Caron & MacLean, 1988; Soken & Pick, 1999). Recently Vaish and Striano (2004), studying the social referencing – a ToM precursor (see paragraph 1.3.1.) – found that voice, even without a visual reference, is more potent than facial cues in guiding infants' behavior. Perception researches have shown that human beings are able to identify emotions from verbal intonation differences in speech (Bachorowski & Owren, 1995; Banse & Scherer, 1996) and that listeners find auditory cues (such as tone, pauses, hesitations) more informative (about a speaker's intention to lie) than visual information from the face (Wiseman, 1995).

For example adults rely to paralinguistic features (intonation) of the verbalization to label the speaker's emotional state, when he presents contradictory emotional information (e.g., happy situation is described with sad paralanguage), like older children, that increase this ability from 5 to 10 years old; on the contrary younger children (4 years old) rely mostly to the linguistic content (Morton & Trehub, 2001).

The auditory channel is therefore apt to be especially effective in emotional communication (e.g., Sorce, Emde, Campos & Klinnert, 1985; Baldwin & Moses, 1996); linguistic and paralinguistic features of emotional speech are processed separately in the brain: verbal content is processed in the left hemisphere, while intonation is processed in the right one (McNeely & Parlow, 2001; Wildgruber et al., 2005).

The sensitivity to cues is a highly adaptive skill and, as suggested Vaish and Striano (2004, p. 266), and it deserves much more attention in the ontogeny of human social cognition. I could add that it deserves also attention in the following developments of mentalization, in fact there are not ToM researches that use this specific channel in older children, from the typical population.

Only some studies in adults with and without autism spectrum conditions, conducted mainly by Baron-Cohen and his team (see the descriptions in the next paragraph), used vocal cues, alone or with facial expressions, to assess ToM and to create training programmes to improve the mind-reading ability.

In this chapter I present a new ToM instrument, called *Voice Test*, that assesses children ability to recognize a wide range of mental states from vocal cues. First I describe why and how the Voice Test was created; then the works of validation and standardization of the instrument on a typical Italian school age children population; finally I suggest some possible uses of it and future directions about the assessment of ToM in children.

2.1. Introduction: Why to create a new instrument?

ToM is a complex, developmental phenomenon, which implies certainly more than just the understanding of false belief. There is a need for assessment tools that measure the developmental progression of ToM in a broader age range. Nevertheless a paucity of studies has explored the development of mentalization ability in school age children belonging to a normal population. Referring to the emotion development, a similar criticism has recently raised also within the field of study of emotion understanding (e.g., Pons, Harris, 2005) and emotion recognition from nonverbal cues (e.g., Rothman & Nowicki, 2004; Tonks et al., 2007). I guess because on the one hand (theoretical point of view) such ability has traditionally been thought to develop in early childhood and remain relatively stable throughout the rest of childhood and adult life (Battacchi, 2004; Nadel & Muir, 2005). On the other hand (methodological point of view) ToM assessment become more difficult in older children and it is hard to find instruments that can investigate the complexity implied in the everyday mental states understanding and in its link to other variables (such as other cognitive and emotional skills and personality aspects).

In these last years there is a increasing trend to find new instruments, modifying tasks already existent or creating new ones, able to assess ToM both in a normal and clinical population. These tasks are called "advanced" or "high-level" tasks (e.g., Happè, 1994; Baron-Cohen et al., 1997; Jolliffe & Baron-Cohen, 1999; Kleinman, Marciano & Ault, 2001; Brent et al., 2004), because they assess the recognition of complex mental states, that classical tests cannot grasp.

There are two main typologies of ToM advanced tasks (Liverta Sempio et al., 2006): tests based on perceptive cues and narrative tasks. This last typology of test assesses the ability to infer epistemic (e.g., irony, white lie, persuasion) and emotional (e.g., moral and mixed emotion) mental states from short stories, usually presented with pictures that facilitate children's comprehension of the plot. Narrative tasks are, for example, the "*Strange Stories*" by Happè (1994; Italian translation: Mazzola & Camaioni, 2002), the "*Faux Pas*" by Baron-Cohen, O'Riordan, Stone, Jones and Plaisted (1999; Italian translation: Liverta Sempio, Marchetti & Lecciso, 2005b), the "*TOM Test*" by Muris and colleagues (1999) and the "*Test of Emotion Comprehension (TEC)*" by Pons and Harris (2000; Italian standardization: Albanese & Molina, in press).

Briefly, the "*Strange Stories*" test consists of 24 short stories about everyday social communication abilities, implying a mentalistic interpretation of behaviors (such as pretence, joke, lie, white lie, figure of speech, misunderstanding, persuasion, sarcasm, double bluff, contrary emotions, appearance/reality and forgetting); children are asked to explain why the story's character acts like that; they usually pass this test around 8-9 years old.

The "*Faux Pas*" consists of 10 stories in which a character has a socially inappropriate verbal behavior (called *faux pas*); children judge if someone in the story said something he/she shouldn't have said, because it would be hurtful; they correctly understand *faux pas* around 11 years old. The "*TOM Test*", used with children between 5 and 12 years of age, comprises 78 questions about stories' characters, subdivided in three subscales of growing difficulty: precursors of ToM (e.g., recognition of emotions, pretence), first manifestations of what the Authors called a "real ToM" (e.g., first order false belief understanding) and more advanced aspects of ToM (e.g., second order false belief understanding, comprehension of humour).

Finally, the "*TEC*" assesses, in 3-11 years old children, specific components of the understanding of emotion, hierarchically organised in three groups in function of their level of difficulty: the first group of components focuses on external aspects of emotions (e.g., recognition of facial expressions, understanding the impact of situational causes on emotions); the second one regards various mental aspects of emotions (e.g., the understanding of the role of beliefs and desires on emotions); and the third group assesses children's understanding of the way in which an individual can think about an emotionally charged event from more than one perspective (e.g., understanding of mixed feelings, cognitive control strategies).

The tests I mentioned up to now imply the use of the channel I called before the cognitive one, in fact characters' mental states are inferred only from events understanding (antecedents, behaviors and statements, described in the story). Children have not information about characters' facial expressions or tone of voice, so these tasks do not require to refer to the perceptive channels to infer mental states. In our daily relational life it is very unlikely a lack of perceptive cues, moreover visual (such as facial expression, gaze, gestures, postures) and auditory (such as tone and rhythm of voice) cues are extraordinary informative signals of mental states. These cues are parts of human nonverbal behaviors (e.g., Darwin, 1872; Ekman & Friesen, 1969; Hinde, 1979) and their codification is important for social relationships (Ricci Bitti & Zani, 2002), in order to give meaning to one's own and others' behaviors.

In the field of ToM studies, in despite of the big amount of data analyzed by general psychology about the nonverbal behavior, only recently the attention has been turned to the ability to infer mental states also from perceptive cues. In the next paragraph I describe various tasks in which visual and/or vocal signals give information about persons' epistemic or emotional mental states.

2.1.1. Theory of Mind tasks based on perceptive cues

This presentation of tasks based on perceptive cues created in the field of ToM follows a chronological order, because it suggests how ToM studies develops, both in theoretical and in methodological background, and it can better explain why there was a need to create a new instrument with vocal cues.

Visual Theory of Mind tasks

One of the first study that used a task based on perceptive cues was developed in 1992 by Baron-Cohen and Cross; using photographs of persons' facial expression, they found that normally developing 4 years old children can infer when someone is thinking from the direction of gaze. In order to analyze deeper this result, Baron-Cohen, Riviere, Fukushima, French, Hadwin, Cross, Bryant and Sotillo (1996) investigated how cognitive mental states (and not only the six basic emotions: happiness, sadness, anger, disgust, fear and surprise[1]) can be inferred from facial expressions.

At the time of their study, they could refer to a lot of works about emotion recognition (e.g., Ekman & Friesen, 1969; Izard, 1971; Zajonc, 1985), while the recognition from face of cognitive mental states were less studied. Only the cognitive states of deception (Ekman & Friesen, 1975; Kleinke, 1986) and interest (Izard, 1971; Ekman & Oster, 1987) focalized some researchers attentions, even within the field of study of emotions.

[1] The authors (Baron-Cohen et al., 1996) agreed with Ekman's (1992, 1999) conception of six basic emotions. But theorists disagree about how many such emotions there are (generally between six and ten) and which emotions are basic (emotion sets include fear, anger, joy, sadness, and disgust; some include surprise, shame, interest, etc.) (e.g., Izard, 1971; Frijda, 1986; see also: Ortony & Turner, 1990; Anolli, 2002; Solomon, 2002), so in the next pages I mention other studies that included a different number of basic emotions.

The aim of Baron-Cohen and colleagues' (1996) study was to investigate whether adults (from a range of social classes, age groups and occupations) are able to perceive a range of mental states and whether this ability is stable across differing cultures. So the tested individuals from European (Britain, Spain) and non-European (Japan) cultures using illustrations of faces took from painting of famous painters; subjects were asked to choose, between two mental words, which best described what the person in the picture was feeling or thinking. They tested 11 different facial expressions (wary, astonished, contempt, recognize, threaten, regret, worried, distrust, revenge, guilt and scheme).

They found that adults have aptitude for reading a wide range of mental states in the face, in a cross-culturally similar way (with the exceptions of wary, guilt and scheme). Finally, Baron-Cohen and colleagues (1996) analyzed if and when children, from 8 to 11 years old, develop the ability to recognize the same set of mental states in the illustrations used in the previous experiments. They did not find age differences and concluded that by the age of 8 children are already very competent at reading the mind in the face, beyond the narrow set of basic emotions.

In the following years Baron-Cohen with his colleagues, within the Autism Research Centre (University of Cambridge), created new instruments based on perceptive cues, searching more subtle tests of mindreading especially to assess ToM in adult with autism, because he found that false belief tasks don't accurately measure the social deficits of this group (in fact some adults with autism tend to pass first and second order false belief tasks, but this result cannot lead to the conclusion that they are necessarily normal in this domain; e.g., Ozonoff, Pennington & Rogers, 1991; Bowler, 1992).

The theoretical root of Baron-Cohen's studies (that we already can glimpse in his previous works and that was made clear in 1997 by Baron-Cohen, Wheelwright and Jolliffe) is the importance he gives to the ability to perceive a range of mental states, both cognitive and affective[2], in interpreting and predicting the actions of others.

In 1997 he published two papers in which ToM were assessed using visual cues. The work by Baron-Cohen, Wheelwright and Jolliffe (1997) used photographs of an actress posing 10 basic emotions (happy, sad, angry, etc.) and 10 complex mental states (admire, interest, thoughtfulness, etc.), testing the role of face parts (eyes, mouth or whole face) in mindreading.

They found that for the basic emotions the whole face is more informative than single parts, while for the complex mental states, seeing the eyes alone produced the best performance; furthermore they pointed out that adults with autism or Asperger syndrome are impaired on the complex mental states recognition, in particular these subjects do not draw information from eyes, they have not a "language of the eyes".

Probably, I suggest, it was the second work published in the same year by Baron-Cohen, Jolliffe, Mortimore and Robertson (1997) that influenced more deeply following ToM researches. In this paper they illustrated a very innovative ToM task for normal adults and adults with high-functioning autism and Asperger syndrome, called the "*Reading the Mind in the Eyes*", or the "*Eyes Test*" for short, which items were first described by Baron-Cohen (1995), then revised in 2001a by Baron-Cohen,

[2] Baron-Cohen and colleagues (e.g., Baron-Cohen, Wheelwright & Jolliffe, 1997; Baron-Cohen, Jolliffe, Mortimore & Robertson, 1997) gives the same meaning to the words "emotional" and "affective" mental states, so – as he does – I use in this paragraph the two terms interchangeably, even if I am aware (and I agree) that some authors, within the psychodynamic view (e.g., Fonagy et al., 2002), use differently these terms. In the next sections when I write about emotions I refer to discrete mental states, temporally and spatially defined, elicited from physics or psychological situation (Ekman, 1992; Ekman & Davidson, 1994), while the term "affective" denotes mental states that characterised all relationships, primarily caregiver-child interaction (Fonagy et al., 2002).

Wheelwright, Hill, Raste an Plumb (Italian normative data: Serafin & Surian, 2004). This final version of the test consists of 36 photographs (in the 1997 version there were 25 photographs) showing persons' (male and female) eye region of the face. Participants are asked to make a forced choice, among which of four mental words (in the previous version of the test between 2 words) best describes what the person might be thinking or feeling. It consists also of a control tasks, the *Gender Recognition Task* (participants look at the same sets of eyes, but this time identifying the gender of person in each photograph), that does not imply mindreading; finally the test includes a Glossary, that subjects can read if they were unsure of the words meaning.

The *Eyes Test* was designed to be "a pure theory of mind test, at an advanced level" (Baron-Cohen et al., 1997, p. 816), that is it does not involve executive function (such as inhibition, planning) and central coherence components (there are not contextual information available). It is more than just an emotion perception test, because it includes terms describing also cognitive mental states; target mental states terms are, for example: upset, desire, fantasizing, regretful, doubtful, preoccupied, interested, serious, nervous, hostile, cautious, contemplative, suspicious.

They found that adults with autism/Asperger syndrome are impaired in this ToM test and that females perform significantly better than males (Baron-Cohen et al., 1997; 2001a). The *Eyes Test* can be considered a "real" measure of ToM, because it correlates to the Happè (1994) *Strange Stories*; furthermore it was demonstrated that adults with autism/Asperger syndrome performed as normal adults in the Gender Recognition Task and in an emotion task that assess basic emotion recognition from faces and they have the same level of general intelligence of normal populations: these data mean that the poor performance is not due to other deficits (e.g., in extracting social information from minimal cues, in basic emotion recognition, perceptual and intellectual deficits)

and that the mental states recognition from eyes can be an important aspect of social cognition (Baron-Cohen et al., 1997).

As the Authors suggested, this test is near to real social situations, obviously it simplifies the stimuli, that in everyday life are not static, but proceed rapidly, requiring the ability to grasp immediately (at an automatic and unconscious level) which intention, motives or other mental states they subtend; furthermore it involves the attribution of the relevant mental state (first stage of ToM attribution), but not of the content of it (that is why he/she has that mental state; it is the second stage of ToM attribution) (Baron-Cohen et al., 1997; 2001a).

The Italian version of the *Eyes Test* (named "*Test degli Occhi*"), by Serafin and Surian (2004), based on the reviewed version of the test (Baron-Cohen et al., 2001a), reports normative data for Italian adults. The authors, that administered the test to a normal population from 18 to 93 years old, pointed out two results, that were different from Baron-Cohen and colleagues finding: they did not find a gender differences in the test, but found that older persons have lower performance than younger adults, both in the *Eyes Test* and in the *Gender Recognition Task*.

Baron-Cohen, with Wheelwright, Spong, Scahill and Lawson (2001b) adapted the *Eyes Test* to create a child version of it (Italian translation: Liverta Sempio, Marchetti & Castelli, 2003) (Figure 2.1. shows an item of the *Eyes Test* – child version).

It comprises 28 photographs of the eye region of the face, chosen among those of the adult version of the test; children are asked to pick which of 4 mental words best describes what the person in the photo is thinking or feeling. Also in this version, the task includes the *Gender Recognition Task*, as a control for the non-mentalistic social attribution. It was administered to normal children and children with Asperger syndrome, age 6-10.

They found that children with Asperger syndrome are impaired in ToM and that, in normal population, older children performed better than younger children, revealing a main effect of age; they did not find gender differences.

Figure 2.1. An example of male eyes of the Eyes Test – child version (Baron-Cohen et al., 2001b); word choices were: hate, unkind, worried (correct), bored.

Auditory Theory of Mind tasks

Recently researchers' attention have been also focused on another perceptive channel: the auditory one. Nearly at the same time, two research groups created, independently between them, two advanced ToM instruments, for adults with autism, based on vocal cues.

In 2001, Kleinman, Marciano and Ault created an auditory ToM task, the "*Mental State Voices Task (MSVT)*", consisting of a sentence without inherent emotional content ("The quick brown fox jumped over the lazy dog"), recorded using different intonations, in order to represent six basic emotions (happy, sad, angry, afraid, surprised, disgusted) and six complex mental states (arrogant, guilty, calm, anxious, bored, interested). Participants choose between to adjective which best describe what the actor was feeling or thinking.

They used, like in the *Eyes Test*, a gender control task using the same voices of the MSVT. Their aim was to compare, in adults with autism, performance in an advanced visual mental state attribution (the *Eyes Test* by Baron-Cohen, Wheelwright & Jolliffe, 1997) with that in the MSVT, considered as a parallel auditory task. Their work extended Baron-Cohen, Wheelwright and Jolliffe's (1997) work by showing that the deficit in individuals with high-functioning autism is not limited to the "language of eyes", but it includes also the identification of complex mental states from voice.

They found that participants were significantly more accurate in assessing mental states based on the voices as contrasted with the eyes task, concluding that "for people to assess emotions, disembodied verbal information appears superior to information gleaned from the eyes" (Kleinman, Marciano & Ault, 2001, pp. 33-34); they failed to find a gender differences in both perceptive tasks.

In the same year, Rutherford, Baron-Cohen and Wheelwright (2001) presented a research with the same aim of Kleinman, Marciano and Ault (2001). In order to expand Baron-Cohen and colleagues' (2001a) finding about the adults with autism/Asperger syndrome from the visual domain into the auditory one, they create a new task: the "*Reading the Mind in the Voice Test*", shortly the "*Voice Test*".

It consists of 40 segments of speech, selected from 50 samples of dialogue recorded from dramatic audio books (e.g., "Collie said you were up here"; "Please! We must go"). Subjects choose between two adjectives to best describe the mental state of the speaker (e.g., referring to the two examples cited above, "friendly" as target term and "grateful" as foil; "worried" as target and "insulted" as foil).

Each segment last for approximately 2 seconds with a 3-seconds pause between sentences, during with participants marked their choice.

The *Voice Test* investigates mentalistic attributions, for example target mental states included in the test are: friendly, suspicious, embarrassed, sincere, reassuring, sarcastic, reflective, hopeful, annoyed, resigned, disappointed, worried. The control task consists of the same voices of the test and participants judge the speaker's age (under or over 42).

They found that people with high-functioning autism/Asperger syndrome have a specific deficit making social inferences in auditory domain, too. The *Voice Test*, like the *Eyes Test* (Baron-Cohen et al., 1997; 2001a), do not correlate with IQ: the complex mental states recognition from perceptive cue is independent of intellectual ability.

The Voice Test was recently revised[3] (Golan, Baron-Cohen, Hill & Rutherford, 2007a), because authors found some limitations: it had only two possible answers per question (while the task's sensitivity could be improved by adding foils to each item) and some items could not discriminated autism spectrum condition group from the normal one. So they removed some items and added foils (selected from the emotion taxonomy created by Baron-Cohen, Golan, Wheelwright & Hill, 2004, consisting in 412 mental states, each in one of six developmental levels). The final version of the test consists of 25 item, presented in a random order, using a computer software. It has good reliability and validity, it is harder and more sensitive than the first version (Rutherford et al., 2002) in distinguishing adults with autism/Asperger syndrome from normal adults.

They also found that the verbal IQ has a significant effect on the task scores and they explain this result referring to the difficulty to choose among several foils, that probably require the involvement of verbal ability to distinguish several potentially correct mental states' verbal labels in order to find the most suitable one.

[3] The revision of the Voice Test (Golan et al., 2007) was published after the present research was done, so the Italian child version of the Voice Test was created independently from the recent English revision.

They did not find differences between male and female in normal adult population and found a weak, even if significant, correlation between the *Voice Test* and the *Eyes Test*, in fact the two tests are related to different perceptual modalities and require the recognition of quite different emotions and mental states (Golan et al., 2007).

Visual and auditory Theory of Mind tasks

It could be useful, as already noted Kleinman, Marciano and Ault (2001), to determine how the joint contribution of facial and verbal information together affects accurate processing of mental states.

Now I briefly describe two new tests, worked out in 2007, that use both visual and auditory channel, getting closer to real social situations, in which rarely faces and voices are separated. These tests were administered to individuals with autism spectrum conditions, in order to analyze deeper their ToM deficit.

Golan, Baron-Cohen and Hill (2006a) constructed a battery of visual and complex emotion recognition tasks, called the "*Cambridge Mindreading Face-Voice Battery*" (CAM), that evaluates, separately, the recognition in the face (Figure 2.2.) and in the voice of 20 emotional concepts, each expressed by 5 items (2 or 3 faces and 2 or 3 voices) from the taxonomy of emotion (Baron-Cohen, Golan, Wheelwright & Hill, 2004). After watching a clip or listening a voice, the participant chooses the word, among 4 adjectives, which better describe how the person is feeling. They showed that the vocal and the visual scales strongly correlated and that males with autism spectrum conditions had more difficult to recognize emotions from faces than from voices (while this difference was not found in females and in the control group) This battery has the merit to test the same emotions, using different perceptive channels.

Golan, Baron-Cohen, Hill & Golan (2006b) created a complex emotion and mental states recognition task, called the "*Reading the Mind in Films*" task (RMF).

It consists of 22 short social scenes, taken from films, including visual input (facial expression, body language, action), auditory input (prosody, verbal content) and context (e.g., a scene taken from "The Turn of the Screw" (1999): a young woman complimenting an older woman on the way she educated the children; the older woman thanks her calmly, then runs towards her with tears in her eyes saying: "Oh miss, I'm so glad you're here"; target term is "overcome", foils are "sociable", "admiring", "liked"). It requires the integration of multimodal information. Participants were asked to identify the protagonist and to label his/her emotion or mental state at the end of the scene, choosing it among 4 words. They found that individuals with autism spectrum conditions performed lower than controls and that the task correlated with verbal IQ, suggesting that participants used verbal content to pick up the protagonists' mental states.

The authors used also the various perceptive tasks to teach individual with autism spectrum condition to improve their mental states recognition (e.g., Golan & Baron-Cohen, 2006).

Figure 2.2. Example of item of the emotion recognition in the face task (Golan, Baron-Cohen & Hill, 2006a); word choices were: restless, sadistic, cherishing, flattering.

Some final suggestions

In conclusion of this historical *excursus* within ToM studies, I briefly underline some aspects. From the methodological point of view, it is evident the growing complexity, during the years, of instrument for perceptual ToM assessment: from pictures to real faces, from static stimuli to dynamic stimuli, from the separation to the integration of multimodal information. Tasks become therefore more and more ecologic, near to real daily life.

A limit of the works I presented regards the population they were created for; in fact all the studies were aimed to better understand the autistic population and not the normal one (even if normal population were tested as control group), so their difficulty was calibrated on a pathological population, that studies revealed to have difficult to understand mental states (e.g., in classical false belief tests, in narrative advanced tasks and in tasks about ToM precursors) (e.g., Happè, 1994; Baron-Cohen, 1995; Surian, 2002; Frith, 2003; Tager-Flusberg, 2007). So items that constitutes these tasks are easier for normal population than for clinical one.

Furthermore they were administered usually on adults, only few tasks were created specifically for the children population, so they cannot analyze the developmental course of ToM after the false belief comprehension.

From the theoretical point of view, I suggest that the creation of these tasks testifies the idea that ToM is a continuous and multi-faceted ability and persons can have different levels in various aspects of the mindreading ability. The so called normal population is able to use, automatically and contemporaneously, different channels to understand mental states and probably they can provide with some high ToM abilities for a possible low level in other specific ToM abilities, so they appear competent in social relationships.

In the developmental context, this ToM view lead researchers to study which and when specific mentalization abilities develop and how (which communicative channel) children grasp mental states during daily life.

The *Voice Test* for the Italian normal children population was created with the aim to analyze deeper ToM (understood as a multi-componential ability, that encompasses a wide range of epistemic and emotional mental states) development in school age children. In fact Italian ToM researchers can count on the translation (but not Italian validation and standardization) of some advanced ToM tasks, regarding the cognitive and the visual channel, but not the auditory one.

The *Voice Test* child version I present in the next section is a new test, developed in Italian (it is not a simply translation of an English task) (phase 1), validated (phase 2) and standardized (phase 3) in a normal children population, from 7 to 11 years old.

2.2. Phase 1: Construction of the Voice Test

The *Voice Test* was developed following three main steps. First: the creation of the items and the preparation of the auditory stimuli. Second: the reduction of items on the basis of the correspondence between auditory stimuli and mental state terms and of the semantic neutrality; then it was run the analysis of three different administration procedures. Third: the analysis of items difficulty and discrimination and the internal consistency of the test and the consequent further items reduction, in order to produce the *Voice Test* final version.

2.2.1. Preparation of the stimuli

On the variety of ways to approach construction of the *Voice Test* items to understand mental states in the voices, the one adopted here was to use adult actors to produce stimulus items. This test procedure, used successfully in previous works within the field of study of the emotion paralinguistic cues (see, for example, Baum & Nowicki, 1998; Rothman & Nowicki, 2004), has the advantage that the production of items was guided by researchers' direction. In fact, initially 48 items were created; each item was constituted of a segment of dialogue, recorded by professional actors and actress and of four terms regarding complex epistemic or emotional mental states, written on the answer sheet. Children were asked to pick which of the four words best describes what the speaker's mental state (only one term, among the four words, correctly represents what the speaker is thinking or feeling). The four options were developed on the basis of the other ToM perceptive tasks: the *Reading the Mind in the Voice* test (Rutherford, Baron-Cohen & Wheelwright, 2002) for adult and the *Eyes Test* – Child version (Baron-Cohen et al., 2001b) translated in Italian (Liverta Sempio, Marchetti & Castelli, 2003). Each set of complex mental terms, referred to epistemic and emotional mental states, was built considering the school age children's social experience. In particular, the terms that in the *Voice Test* by Rutherford, Baron-Cohen and Wheelwright (2002) were typical of adult life (e.g., flirtatious) were not included in the test. Besides mental terms taken from the *Voice Test* and the *Eyes Test* – Child version, it were included others mental states typical of children experiences (e.g., pleading). So there were 48 target mental states, each associated to three foils.

The segments of dialogue of each item were identical to those, considered suitable for children, used in the *Reading the Mind in the Voice* task (Rutherford, Baron-Cohen &

Wheelwright, 2002) or were specially created for the present task, following the structure of Rutherford, Baron-Cohen and Wheelwright's (2002) sentences. All the segments of dialogue were semantically neutral, in fact they can be spoken in a number of different ways to communicate a specific mental states at various level of intensity. So each item refers to an epistemic or emotional mental state depending on the way it is pronounced by the actor/actress (non verbal cues), independently from the semantic content of the phrase (for example, the segment "I know you met Tina" is pronounced in a particular way – using a specific tone of voice, rhythm… – associated to the target mental term "thinking to something").

A Glossary with explanations and examples for all the mental words were built, that participants can look it up if they need an explanation of mental terms. The Glossary was built following the glossaries created for the other ToM tasks based on perceptive cues, from which the mental states were taken (i.e., the *Reading the mind in the Voice* and the *Eyes Test* – Child version) (see Appendix).

The translation process of the mental state terms, the spoken phrases from the adult version of the Voice Test (Rutherford, Baron-Cohen & Wheelwright, 2002) and the Glossary followed the standard guidelines (Hambleton, 1994); two Italian translators, perfect English speakers and expert in the field of ToM, translated the items, then an English native speaker back-translated them and compared her version to the original version.

Stimuli were recorded using the voices of two females (one under 25 and one over 60 years) and two male (one under 25 and one over 60 years) Italian adults, with current theatrical acting experience. Sentences were recorded on an audio cassette. Each speech segment lasted for approximately 2 second, with a 3-second pause between speech segments.

This test requires to choose the correct answer, among four options, while the English *Voice Test* (Rutherford, Baron-Cohen & Wheelwright, 2002) involved a forced choice between only two response options (but the chance performance on each trial was high: $p = 0.5$). The task's sensitivity (as noted also recently Golan, Baron-Cohen, Hill & Rutherford, 2007) could have been improved by adding foils, so the Italian *Voice Test* had two more foils for each item (similarly to the revised version of the *Eyes Test* by Baron-Cohen et al., 2001a, and the child version of the *Eyes Test* by Baron-Cohen et al., 2001b). Moreover, taking (target and foils) mental states from existing validated ToM tests satisfied the content validity of this new test: the items were representative of the mental states the test was originally designed to measure (Boncori, 2006).

For control trials, the answer sheet had two options to represent the person's gender ("male" and "female") and two option to represent the person's age ("under 25" and "over 60"). The original English version of the test asked only the person's age, the gender recognition task was here added because children, having less life experience than adults, can find difficult to recognize persons' age (for example, to recognize the possible change in voice features during the course of life) and in order to make similar the child version of both the *Voice Test* and the *Eyes Test*. The entire procedure I describe in this paragraph, to conceive the *Voice Test* stimuli, was developed by O. Liverta Sempio, and A. Marchetti of the Theory of Mind Research Unit, Catholic University of the Sacred Heart, Milan.

2.2.2. Item reduction and administration procedure

Four independent adult judges were asked to judge the speaker's mental state, among the four options, as they listened to the audiotape; they took the test in real time, with

a 3-second pause between segments. Seven items on which the judges were not unanimous were kept, leaving 41 segments of speech. The same procedure was followed for the control task: again, four independent judges were given the questionnaire and asked to judge speaker's age and gender, between the two options, as they listened to the audiotape; all the judges answered correctly to this task (Liverta Sempio, Marchetti & Fabio, 2005).

To verify if the 41 items were neutral in their meaning, 22 children (13 male and 8 female, coming from a middle-class background, which parents gave their consent to participate to this research task) without learning problem, attending the third class in a primary school near Milan, were asked to judge if the sentences expressed an intense negative or positive meaning, or a neutral, or a very little negative and positive meaning. This procedure selected sentences which meaning did not influence children's mental state attribution to the speaker, in order to make the *Voice Test* a task that assesses the ability to understanding mental states only from nonverbal vocal cues (paralinguistic signals).

Each child read by him/herself the sentences written on the answer sheet; 4 control sentences (with a very negative and positive meaning) were added to the 41 items and they were correctly judged from all the children. Four target sentences did not receive a judgment of neutrality (including the judgment of neutral, very little negative and very little positive meaning) from the 75% of children and were removed from the test, leaving 37 items.

Before the Voice Test creation proceeded any further, it was verified if the test could be administered, without altering children's performance, also in group. This step was important because a group setting reduce the time for administration (more data collection in less time) and allows to use easily the test in different contexts (for

example not only experimental one, but also school context). To this aim children were tested in different settings.

Participants

Sixthy-six children (33 males and 33 females), born in 1998 (mean age: 88,20 months; standard deviation: 3,30 months), were recruited from four primary schools located near Milan; they all attended the second class and came from a middle-class background. Two subjects (a male and a female) were removed because they had learning problems.

All children's parents gave their written consent to allow children's participation in this study.

Materials and Procedures

Children were administered the 37 items of the Voice Test (preceded by the trial item), in a quiet room at their school; they were asked to pick which of 4 words best describes what the speaker is thinking or feeling. Immediately before listening each items, participants were shown and read the 4 options on the answer sheet and asked whether they did not understand any words or they were unsure of any words meaning. If they needed an explanation of a word, the experimenter read aloud some synonymous and an example from the Glossary. The experimenter paused the audiotape between speech segments to give children the time they needed to choose the mental state term.

After the administration of the test, they listened again to the same 37 items (preceded by the trial item) and they were given the control task (recognition of the speakers' gender and age): they marked if the speaker was a male or a female and if he or she was

young (under 25 years) or old (over 60 year). Children were allowed to ask to adjust the volume as necessary before and during the test.

The same procedure (see Appendix for detailed instruction) was also used in the following administration.

Participants were tested in three different setting:

- 20 children (10 males and 10 females) were tested individually;
- 24 children (12 males and 12 females) were tested in little groups of three children (four groups made up of two males and a female and four groups of two females and a male);
- 20 children (10 males and 10 females) were tested in group of 10 children (5 males and 5 females in each group).

Results

An analysis of variance was conducted to compare children's performance in the three administration settings. The one-way ANOVA with the condition of administration as the independent variable and children performance in the Voice Test as dependent variable did not show significant differences between means for different setting: $F(2, 61) = 1{,}123$, $p = 0{,}33$.

All the children passed the gender recognition task. To analyze if children were able to recognize the speakers' age, a frequency analysis was run for each item. Thirty-one items received by the 40-60% of the participants the correct answer, only six items were correctly recognized by most of the children (>75%).

An ANOVA was run to analyze if the administration setting influenced the recognition of the speakers' age, but it did not reveal significant differences: $F(2,61) = 0{,}565$, $p = 0{,}571$.

Discussion

The performance in the Voice Test in the three different administration setting (individual, little group of three children, group of ten children) was not different; the following administration therefore could be done in group of 10 children.

About half of the children failed in the age recognition control task and this result was not due either to the administration context or to a lack of attention or auditive problems, because they correctly identified the speakers' gender. So this task was removed from the test because of its difficulty for children and only the gender recognition task remained as a control task, like in the *Eyes Test*.

2.2.3. Further item reduction: the final version of the Voice Test

The first step was to sample the 37 items onto a computer and digitally clean tape recording noise. The task was therefore presented, from now on, to the participants using a CD player. Moreover the answer sheet for control trials was changed in order to represent only the two gender options ("M" for male and "F" for female).

In order to create a valid test, it was necessary to identify the items that were too much simple or difficult; to analyze internal consistency of the *Voice Test*; and to verify if children's accuracy in identifying the mental state in voices increased with age.

Participants

170 children (80 male and 90 female; mean age: 100,09 months; s.d.: 16,59 months), aged between 73 and 131 months, were recruited from five primary school located near Milan. They received their parent's written consent to participate in the study. They were all Italian native speaker and came from a middle-class background.

They were subdivided in 5 groups, corresponding to the class they were attended:

- first class: 29 children (14 males and 15 females) from 73 to 84 months; mean age: 77,34 months (s.d..: 2,98);
- second class: 32 children (16 males and 16 females) from 83 to 93 months; mean age: 87,95 months (s.d.: 3,41);
- third class: 38 children (13 males and 25 females) from 97 to 108 months; mean age: 101,58 months (s.d.: 3,40);
- fourth class: 38 children (18 males and 20 females) from 108 to 120 months; mean age: 113,61 months (s.d.: 3,78);
- fifth class: 33 children (19 males and 14 females) from 120 to 131 months; mean age: 125,85 (s.d.: 3,75).

Materials and Procedures

Children were tested in groups consisting of a maximum of 10 children (children attending the first class were tested in group of three, in order to better control their understanding of the test procedure), in a quiet room in their school.

They listened to the voices, recorded on the CD, and were asked to make a judgment on each speech segment, choosing a mental state term among 4 words. The *Voice Test* answer sheet and the instruction were the same written in the previous section (see also Appendix); the answer sheet of the control trials required only to mark if the speakers were male or female, so children were asked to make gender judgments for the same segments on which they had made mentalistic attribution before. While in the Voice Test, the experimenter paused the CD-tape between segments (without repletion of the items), in the gender recognition task children took the test in real time, in fact,

according to the experimenter observation during previous administration, 3 seconds was sufficient time for every subject. The Voice Test and the control task took about 20 minutes.

Results

All the children answered correctly to the gender control task.

The statistics typically used in the item analysis of tests are those referred to item facility/difficulty, internal consistency of the test and item discrimination.

The index of difficulty was calculated for all the individual items using the Guilford's (1956) difficulty index: $Pc = (np - 1) / (n - 1)$ [4]. Pc is the corrected proportion (it corrects the formula of p, that is the proportion of children who answered the item correctly) and it can take a score included between 0 (very facile) and 1 (very difficult). In this work it was chosen an intermediate level of difficulty: Pc = 0,50 (+/- 0,20, that is $0,30 < Pc < 0,70$).

The reason for measuring item difficulty is to choose items of suitable difficulty level which help in assessing as accurately as possible each individuals level of knowledge. Thus an item with an index of difficulty higher than 0.70 and lower than 0,30 is deemed to be a poor discriminator.

Guilford's indexes for each item are presented in Table 2.1. (see also Chart 2.1.). The items did not satisfy the intermediate level of difficulty were 15 and they were removed from the *Voice Test*; the test was therefore constituted of 22 items.

[4] In the Guilford's formula "n" represents the number of options, in the case of Voice Test it was 4; "p" is the proportion of correct answers ("number of correct responses / total number of response").

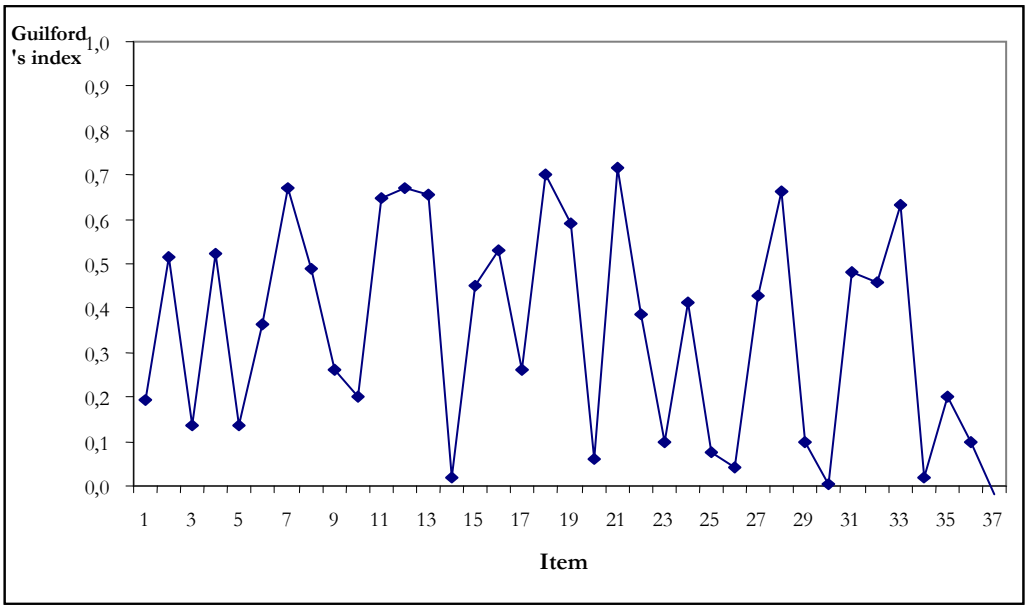
Table 2.1. Guilford's difficulty index (Pc) for each item (selected items are highlighted)

Item	*1*	***2***	*3*	***4***	*5*	***6***	***7***	***8***	***9***	*10*	***11***	***12***	***13***
Pc	0,2	**0,5**	0,1	**0,5**	0,1	**0,4**	**0,7**	**0,5**	**0,3**	0,2	**0,6**	**0,7**	**0,7**

Item	*14*	***15***	***16***	***17***	***18***	***19***	*20*	***21***	***22***	*23*	***24***	*25*	*26*
Pc	0,0	**0,5**	**0,5**	**0,3**	**0,7**	**0,6**	0,1	**0,7**	**0,4**	0,1	**0,4**	0,1	0,0

Item	***27***	***28***	*29*	*30*	***31***	***32***	***33***	*34*	*35*	*36*	*37*
Pc	**0,4**	**0,7**	0,1	0,0	**0,4**	**0,4**	**0,6**	0,0	0,2	0,1	0,0

Chart 2.1. Guilford's difficulty indexes



The determination of the internal consistency for the 22 items of the Voice Test was accomplished through the Cronbach's coefficient *alpha*. Cronbach's alpha was 0,60; in order to improve alpha coefficient two item were removed, because they were poor contributors to the overall measure. Internal consistency for the 20 items was 0,69.

Finally, it was analyzed if *Voice Test* differentiates correctly among different ages. The total score was distributed normally (Chart 2.2.), with a mean of 13,18 (s.d: 3,27), a median of 14,00 and a mode of 12,00. The mean of the total error score was 6,82 (s.d.: 3,27). The correlation between chronological age and total number of errors on the test for all participants was significant: $r = -0,541$, $p < 0,001$.

Table 2.2. shows means and standard deviations of the total error scores and total scores of participants, subdivided in five groups of age; Chart 2.3. represents the *Voice Test* total score increase with age.

An ANOVA was run, with age (the five class) as independent variable and the Voice Test total score as dependent variable: $F(1, 4) = 19,99$, $p < 0,001$, $\eta^2 = 0,33$.

Table 2.2. Total error score and total score of the Voice Test

Mean age (months)	*Total error mean (s.d.)*	*Total score mean (s.d.)*
77,34 (1st class)	8,97 (3,05)	11,03 (3,05)
87,95 (2nd class)	8,81 (3,22)	11,19 (3,23)
101,58 (3rd class)	7,45 (2,46)	12,55 (2,46)
113,61 (4th class)	4,84 (2,16)	15,16 (2,16)
125,85 (5th class)	4,58 (2,72)	15,42 (2,73)

Chart 2.2. The normal distribution of the Voice Test total score

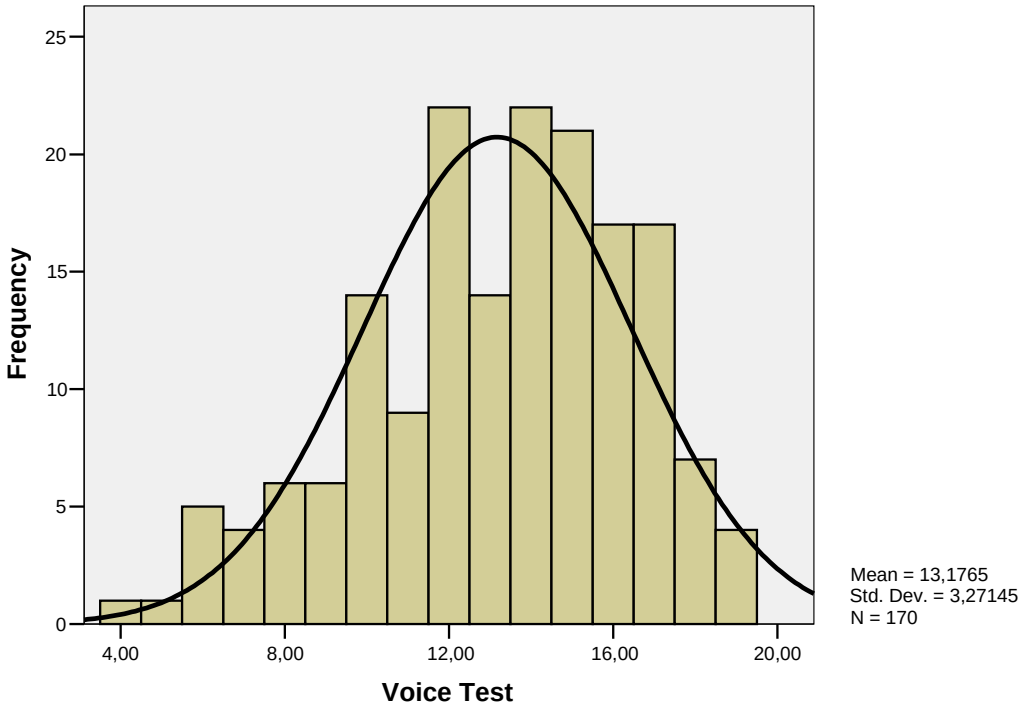
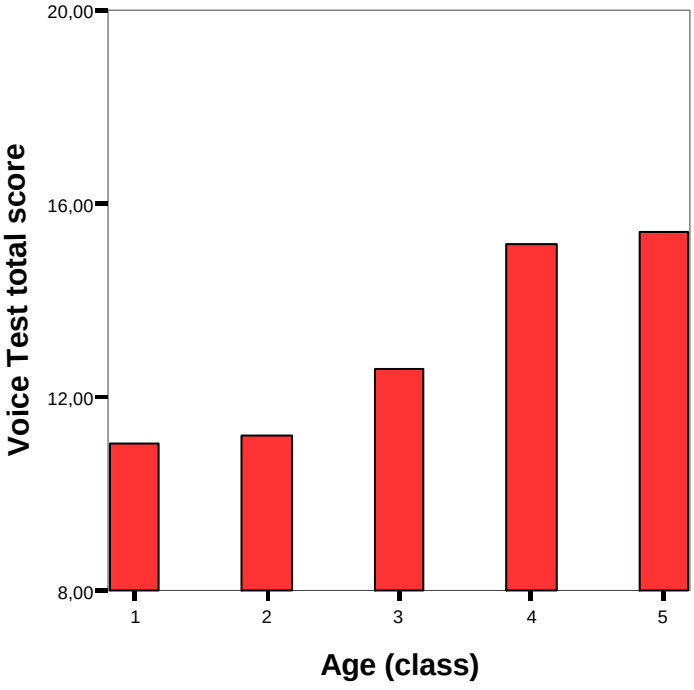


Chart 2.3. The Voice Test score increments with age



Discussion

Briefly, preliminary items analyses were performed in order to obtain the initial validation of the *Voice Test*. The final version of the test consisted of 20 item (see Appendix); 17 items were removed after the analyses of item difficulty (Guilford index) and of the internal consistency. The test showed an acceptable estimate of internal consistency (Cronbach's *alpha* was 0,69) (Cohen, 1960).

It was also verified that the test scores were normally distributed and that children's performance increased with age, from 6,4 to 10,5 years. The results indicated that the psychometric qualities of the Voice Test final version appear to be sound.

2.2.4 Conclusion

The current study describes the three steps of the development of the *Voice Test*, an advanced Theory of Mind test based on auditory cues.

From the 48 initial items, consisted of speech segments representing specific complex mental states acted by two Italian actors and two Italian actress, were excluded 28 items, in order to create a quick, easy to use and valid test.

A series of analyses were run in order to verify: the correspondence between auditory stimuli and mental state terms; the speech segments semantic neutrality; the suitability of the group administration procedure and of the control tasks; the items difficulty; the items discrimination among five groups of age (within children attending the primary school, from the first to the fifth class); the normal distribution of the *Voice Test* total score; and the internal consistency of the test.

The final version of the *Voice Test* consists of 20 items (preceded by the trial item), of which 10 acted by two female voices and 10 by two male voices, representing both

epistemic and emotional complex mental states, followed by a control task implying a social but non-mentalistic ability (gender recognition task). It also includes a Glossary that helps children to better understand mental state terms meanings.

Children were asked, for each item they listened, to pick which of 4 words best describes what the speaker is thinking or feeling. The *Voice Test*, in its final composition (see Appendix), takes about 20 minutes (30 minutes with younger children and 15 minutes with older children).

The test can be really considered an advanced ToM test for children because it assesses the understanding of a wide range of mental states from vocal cues; for example it goes beyond the concepts of desire and belief including the comprehension of more developed mental processes such as "thoughtful" and "confused", and it assesses the recognition ability of not-basic emotion, such as "nervous" or "worried".

The *Voice Test* is therefore an ecological test that differs both from other ToM tests (e.g., Happè, 1994; Baron-Cohen et al., 2001a, 2001b; Rutherford, Baron-Cohen & Wheelwright, 2002) – because of the stimuli it is constituted (voices) and of the population it refers (normal children) – and from other auditory tests regarding emotional competence (e.g., Rothman & Nowicki, 2004) – because of the wide range of mental states it assesses.

In conclusion, the preliminary analyses demonstrated the acceptability of the administration group (maximum of 10 children) procedure, of the content validity and of the psychometric properties of the *Voice Test*.

The next paragraphs present the validation and standardization of the test in a large children population sample.

2.3. Phase 2: Validation of the Voice Test

Validity refers the extent to which a test measures what it is supposed to measure ("Is the test measuring what you think it is measuring?). It is a judgment made on the basis of experience and empirical indicators (Boncori, 2006). In this work it were analyzed the content validity and the construct validity of the *Voice Test*.

Content validity is a non-statistical type of validity that involves the examination of the test content to determine whether it is a representative sample of the specific domain to be measured. This type of validity id ordinarily to be established deductively (Cronbach & Meehl, 1955), I describe in the paragraph 2.2.1. the content validity of the *Voice Test*.

Construct validity refers to the degree to which inferences can legitimately be made from the operationalizations done in the study to the theoretical construct on which the operationalizations were based (Cronbach & Meehl, 1955; Pedrabissi & Santinello, 1997).

In order to verify this type of validity, it were analyzed the convergent validity of the *Voice Test*, that is the degree to which the *Voice Test* is correlated with other ToM measures (e.g., the *Voice Test* is theoretically predicted to correlate with another ToM perceptive test).

The aim of the present study was to validate the new advanced ToM instrument based on auditory cues, the *Voice Test*, on an Italian sample constituted of normal children attending the primary school.

2.3.1. Method

Participants

The sample consisted of 220 children (111 males and 109 females), aged between 78 and 134 months (M = 105,17 months; s.d. = 16,73), attending the primary school in three towns near Milan. Children came from a middle-class background. They had not psychological or neurological proclaimed pathologies and learning difficulties. A few participants (less than 3%) were not Italian native speakers, but all were competent in speaking and understanding Italian. The participants' parents gave their written consent to allow children's participation in this study.

Participants were subdivided in five groups of age:

- 54 children (25 males and 29 females): 7 years old (78-89 months);
- 41 children (23 males and 18 females): 8 years old (90-101 months);
- 48 children (25 males and 23 females): 9 years old (102-113 months);
- 46 children (27 males and 19 females): 10 years old (114-125 months);
- 31 children (11 males and 20 females): 11 years old (126-137 months).

Measures

The *Voice Test* is an advanced ToM task consisted of 20 auditory stimuli (the test creation is described in the paragraph 2.2.; see also Appendix). It assesses ability to read complex epistemic and emotional mental states from voices. Children were asked to judge what the speaker is thinking or feeling (his/her mental state), from his/her voice alone, choosing a word among four mental state terms; then they judge the same speakers' gender (male or female).

Children were tested on three ToM tasks: two classical first-order (*The Deceptive Box*) and second-order (*Look Prediction*) false belief tasks and an advanced measure of ToM ability, based on perceptive cues, the *Eyes Test* – child version.

The *Deceptive Box Task* (Perner, Leekam & Wimmer, 1987; Italian adaptation: Liverta Sempio & Machetti, 2001c) assesses the first-order false belief. Child was shown a closed box of pastels (it was a well-known brand of pastels) and asked what they thought was inside. The box was then opened to show the child that it contained not pastels, but a little doll. The doll was replaced and the child was then asked a memory control question: "What is inside here?". Then the child was asked to predict what a friend of his/her, who did not look into the box because he/she was not in the room, would think was in the box. After this question, regarding the false-belief of another person, the experimenter asked a question about child's false belief: "Before I opened this box, what did you think was inside?".

The *Look Prediction* is a second-order false belief task (Sullivan, Zaitchik, & Tager-Flusberg, 1994; Italian adaptation: Antonietti, Liverta Sempio, Marchetti & Astington, 1999). Children were told a story, with four pictures helping them to follow the plot, about John and his sister Mary: John and Mary are playing in John's bedroom with a new pack of cards, when Mary tells her brother she is going to put cards into the wardrobe, because she will go to the kitchen to lay the table; when Mary goes out the bedroom, John decides to play a trick to her: he takes the cards from the wardrobe and puts them under his bed. Children were asked two control questions: "Did Mary see John hiding the pack of cards under his bed?" (memory question) and "Where does Mary think is the pack of cards?" (reality question). After passing these questions, they were asked a first-order false belief question ("Does John think that Mary saw him hiding the pack of cards under his bed?") and a second-order false belief question:

"Where does John think that Mary looks for the pack of cards when she comes back into his room? And why does she look for there?".

The answers in these two false belief tasks were scored dichotomously as pass-fail; for every correct answer children got one point, for a maximum of four points in each task.

The *Eyes Test* – child version (Baron-Cohen, Wheelwright, Spong, Scahill & Leason, 2001b; Italian version: Liverta Sempio, Marchetti & Castelli, 2003) assesses the ability to recognize and attribute mental states from visual cues. It consists of 28 photographs of the eye region of the face; children were asked to pick which of four words best describes the person in the photo is thinking or feeling; only a word is correct. Children were also administered a control task, in which the same 28 photos were shown and they were asked to judge the person's gender. For each correct answer, children got one point. Information on reliability and validity is not available for the test, but Baron Cohen and colleagues (2001b) found that the total score improve across the age-groups (6-8; 8-10; and 10-12 years) and ANOVA testing revealed that the *Eyes Test* was highly significant in distinguishing between differentiated performance of these groups. The adult version of the test (both in English and in Italian; Baron-Cohen et al., 2001a; Serafin & Surian, 2004) is validated.

Finally, as verbal measure, it was administered a test of verbal mental age: the *Peabody Picture Vocabulary Test – Revised* (PPVT-R) (Dunn & Dunn, 1981; Italian validation: Stella, Pizzoli & Tressoldi, 2000); it is a measure of receptive vocabulary for standard Italian, it is a norm-referenced power test containing 5 training items followed by 175 test items arranged in order of increasing difficulty.

Procedure

Children were tested over two session spaced on two weeks apart, in a quiet room in their school. At session 1, children were given in group of ten children the *Voice Test* and individually the PPVT-R and the first-order *False Belief* task. At session 2, children were tested individually on the second-order *False Belief* task and the *Eyes Test* – child version.

2.3.2. Results

Descriptive statistics

Table 2.3. presents means and standard deviations for all variables. The first-order *False Belief* task was not passed only by eight children; the first-order question in the second-order *False Belief* task was passed by 213 children, while the second-order question by 178 children; these tasks had not a normal distribution. The first-order false belief task did not included in following analyses, while it was used the non parametric statistics when the second-order *False Belief* task was considered in the analysis. All the other tasks were normally distributed.

All the children participating in this study passed the control questions of the *False Belief* tasks and the *Voice Test*; 115 children recognized correctly the person's gender in the control task of the Eyes Test, however all the children recognized correctly at least the gender of 20 persons out of 28 (mean: 26,61; s.d.: 1,79).

Table 2.3. Means and standard deviations

	M	*s.d.*
Voice Test (0-20)	13,10	3,51
Eyes Test (0-28)	15,85	4,14
First-order False Belief (0-4)	3,95	0,25
Second-order False Belief (0-4)	3,79	0,46
PPVT – R	101,94	15,00

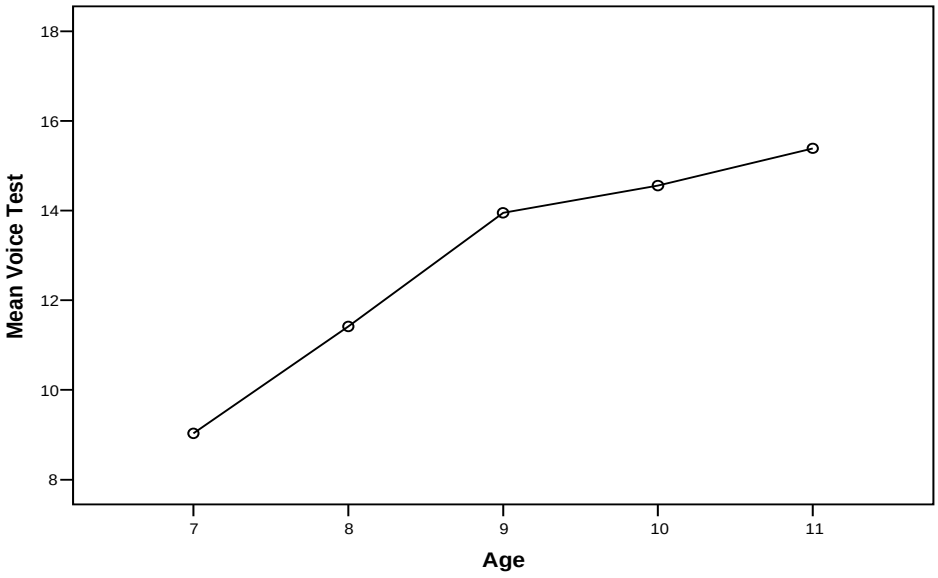
Influence of gender and age

The independent-sample *t* test did not show gender differences in the *Voice Test*, in the *Eyes Test* and in PPVT – R.

The nonparametric test, Mann-Whitney *U* Test, was run to investigate gender differences in the second-order *False Belief* task and showed a significantly difference between male and female (female performance better than male): *U* (218) = 4937,500, p = 0,001, two-tailed.

Both the *Voice Test* and the *Eyes Test* score increased with age: F (4, 215) = 34,075, $p < 0,001$ and F (4, 215) = 15,301, $p < 0,001$. The LSD post-hoc test revealed that between 9 and 10 years old and between 10 and 11 years old, the performance in the *Voice Test* did not increase significantly (Chart 2.4.). The PPVT – R increased with age [F (4, 215)= 8,694, $p < 0,001$], but the LSD post-hoc revealed that differences were only between 6-7 and 7-8 years; there were not differences among 9, 10 and 11 years old. The Mann-Whitney *U* Test was run to analyze if the performance in the second-order *False Belief* increased with age: differences among age were significant ($p < 0,01$), except among 9, 10 and 11 years old.

Chart 2.4. Age influence on the Voice Test performance.



Bivariate correlations

Person's correlation coefficient *r* was used to analyze correlations between normally distributed variables, while its nonparametric version, Spearman's *rho*, was used when the second-order *False Belief* (2nd FB) score was considered. Table 2.4. presents the bivariate correlations between variables. In detail, the *Voice Test* correlated positively with the *Eyes Test* (Chart 2.5.), the second-order *False Belief* task (Chart 2.6.) and the receptive language ability (PPVT-R) (Chart 2.7.).

Controlling for verbal ability the correlation between the *Voice Test* and the *Eyes Test* remained significant: $r = 0{,}404$, $p < 0{,}001$. Also controlling for age it was significant: $r = 0{,}363$, $p < 0{,}001$. Finally, the partial correlation controlling both for age and PPVT – R was significant: $r = 0{,}269$, $p = 0{,}002$.

Controlling for age the correlation between the *Voice Test* and the PPVT – R remained significant: $r = 0{,}329$, $p < 0{,}001$.

Chart 2.5. Correlation between the Voice Test and the Eyes Test

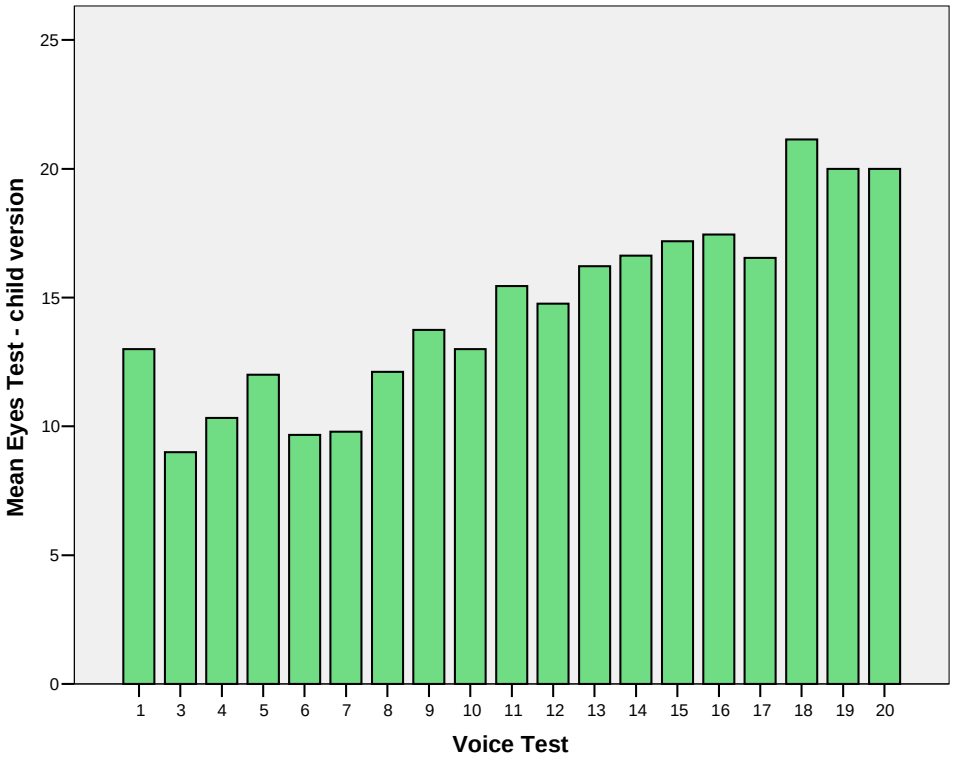


Table 2.4. Correlations (the Voice Test correlations are highlighted)

	Eyes Test	*2nd FB*	*PPVT – R*
Voice Test	0,532 $p < 0,001$	0,357 $p < 0,001$	0,408 $p < 0,001$
Eyes Test	1	0,352 $p < 0,001$	0,422 $p < 0,001$
2nd FB		1	0,144 $p<0,05$

Chart 2.6. Correlation between the 2nd order False Belief task and the Voice Test

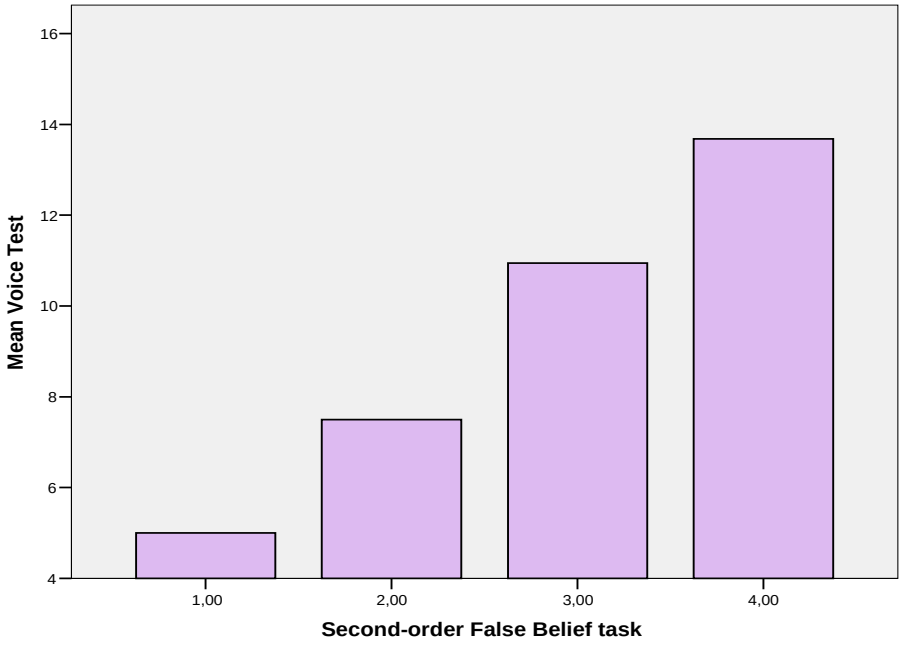
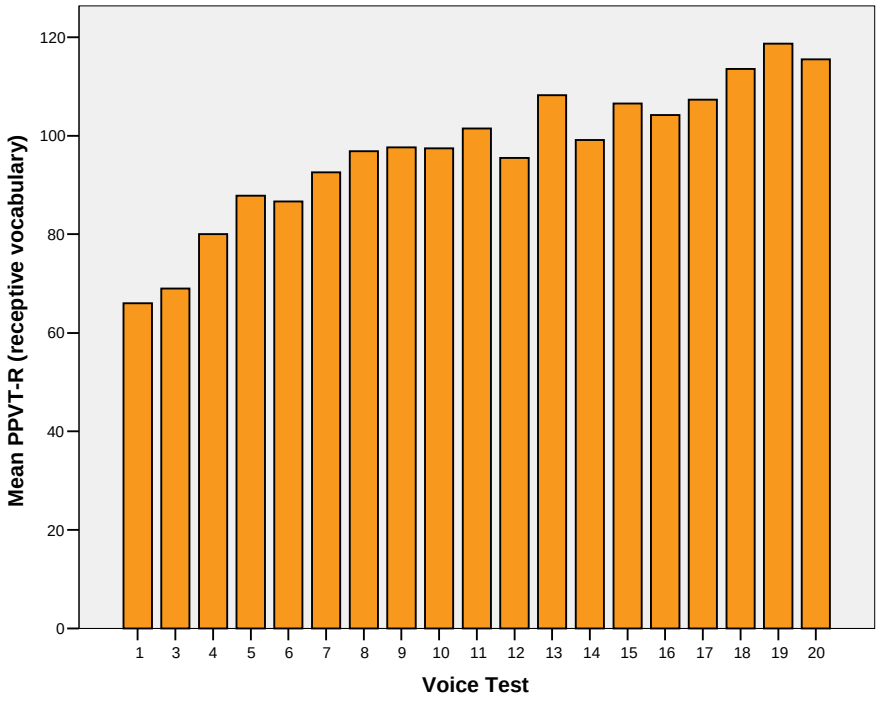


Chart 2.7. Correlation between the Voice Test and the PPVT – R.



2.3.3. Discussion and conclusion

The present study investigated the construct validity of the *Voice Test*, an advanced measure of the mindreading ability for school age children, based on auditory stimuli.

It would expected that the *Voice Test* was related to other Theory of Mind measurements, in particular to those based on perceptive cues, and was not merely a measure of verbal ability (comprehension of the content of the spoken phrases and of the mental state terms). This hypothesis was confirmed by the pattern of correlations of the *Voice Test* with the other variable.

The *Voice Test* and the *Eyes Test* increased with age. As expected, the first-order false belief task was correctly solved by the school age sample, because children understand this type of false belief at around 4 years old (Wimmer & Perner, 1983; Wellman, Cross & Watson, 2001; Flynn, 2006).

Most children also passed the second-order false belief task, in fact it is usually understood around 7 years old (Perner & Wimmer, 1985). There were a gender differences only in the second-order false belief task, as other ToM researches suggested (e.g., Charman, Ruffman & Clements, 2002).

It was found that the *Voice Test* performance correlated positively with other ToM tasks, both classical second-order false belief task and an advanced ToM task based on visual cues, the *Eyes Test*. Correlations remained significant also when verbal ability and age were controlled.

These data confirmed the convergent validity of the *Voice Test*: it assesses the ability to read mental states. Even thought the correlations were high, but the correlation coefficient was not over 0,8 (that would have meant a quite perfect overlapping of the tests); this result can suggest that the construct of Theory of Mind is wide: it comprises

a lot of mental processes, both emotional and cognitive, and it is referred both to the "cognitive" understanding of false belief and to the "visual" and "auditory" communicative channels. However it is enough homogeneous because the different aspects composed Theory of Mind are strictly linked. This connection is more evident in daily life, during interpersonal relationships, where the mental state understanding from vocal cues is interlaced to that guided by visual cues and to the false belief comprehension.

The correlation between the *Voice Test* and a standardized verbal ability measure (PPVT – R), that follows closely the same trend of the other ToM tasks (*False Belief* task and *Eyes Test*), seems to confirm the close relationship and interdependence in development between ToM and language ability (e.g., De Villiers, 2000; Malle, 2002; Astington & Baird, 2005; Antonietti, Liverta Sempio & Marchetti, 2006; Siegal & Varley, 2006), even controlling for age (Milligan, Astington & Dack, 2007).

In conclusion, this study provided support for the construct validity of the *Voice Test*, a new measure of ToM based on perceptive (vocal) cues.

The previous and the present studies showed the psychometric properties of the test (reliability and validity); in the next section I present its standardization.

2.4. Phase 3: Standardization of the Voice Test

A standard test construction procedure requires – after the definition of the construct of interest, its operationalizations into measurable items and the analyses of the psychometric qualities (validity and reliability) – to standardize the observed scores,

indicating how they are distribute with regard to the specific population of people, through the z-scores transformation.

The aim of the present study was to complete the *Voice Test* validation and standardization, identifying school age norms (standard scores for five group of age, from 6,5 to 11,4 years old).

2.4.1. Method

Participants

A total of 586 children participated in the study (the sample included the 220 children that were tested during the second phase of the research; see paragraph 2.3.). There were 295 males and 291 females aged between 78 and 137 months, attending six primary schools located in middle-class neighbourhood of Milan.

They had not psychological or neurological proclaimed pathologies and learning difficulties. A few participants were not Italian native speakers (less than 3%), but all were competent in speaking and understanding Italian. Their parents gave their informed consent to participate in the study.

Children were subdivided in five groups of age:

- 122 children (59 males and 63 females): 7 years old (78-89 months);
- 126 children (68 males and 58 females): 8 years old (90-101 months);
- 136 children (64 males and 72 females): 9 years old (102-113 months);
- 123 children (65 males and 58 females): 10 years old (114-125 months);
- 79 children (39 males and 40 females): 11 years old (126-137 months).

Procedure

Participants were tested in group (consisted of a maximum of 10 children; see paragraph 2.2.2.) in a quiet room at their school. Children were instructed to find and mark, in the answer sheets, the term, among the four mental words written within the same row, that better represented the speakers' mental states (see specific instructions in Appendix).

Participants were said to remain silent and to pay attention to the sentence because they could listen to it once. Their comprehension of the test procedure and the appropriate level of the volume were tested using a trial item.

Before listening each item, were read the 4 options on the answer sheet and asked whether they did not understand any words or they were unsure of any words meaning; if they needed an explanation of a word, the experimenter read aloud some synonymous and an example from the Glossary (Appendix). The experimenter paused the audiotape between speech segments to give children the time they needed to choose the mental state term.

Then, at the end of the test, they received new instructions about the control task; the experimenter showed the answer sheet with two choice ("M" for male and "F" for female) for each item. Children listened again to the same items recorder on the CD (preceded by the trial item) and they marked the speaker's gender in real time.

For each correct answer, children got one point, so the total score for both the *Voice Test* and the control test could vary from 0 to 20. The *Voice Test* took about 20 minutes.

2.4.2. Results

Descriptive statistics

Exploratory analysis suggested the Voice Test score to be normally distributed (Chart 2.8.): mean was 13,23; median and mode coincided and were 14,00; standard deviation was 3,32; skewness and kurtosis of the distribution were respectively -0,787 (s.d.: 0,10) and 0,363 (s.d.: 0,20). The distributions of the Voice Test in each group of age met normally requirements; Table 2.5. reports mean and standard deviation of the Voice Test in each group of age.

All the participants recognized correctly the speakers' gender in the control task.

Chart 2.8. Voice Test total score distribution.

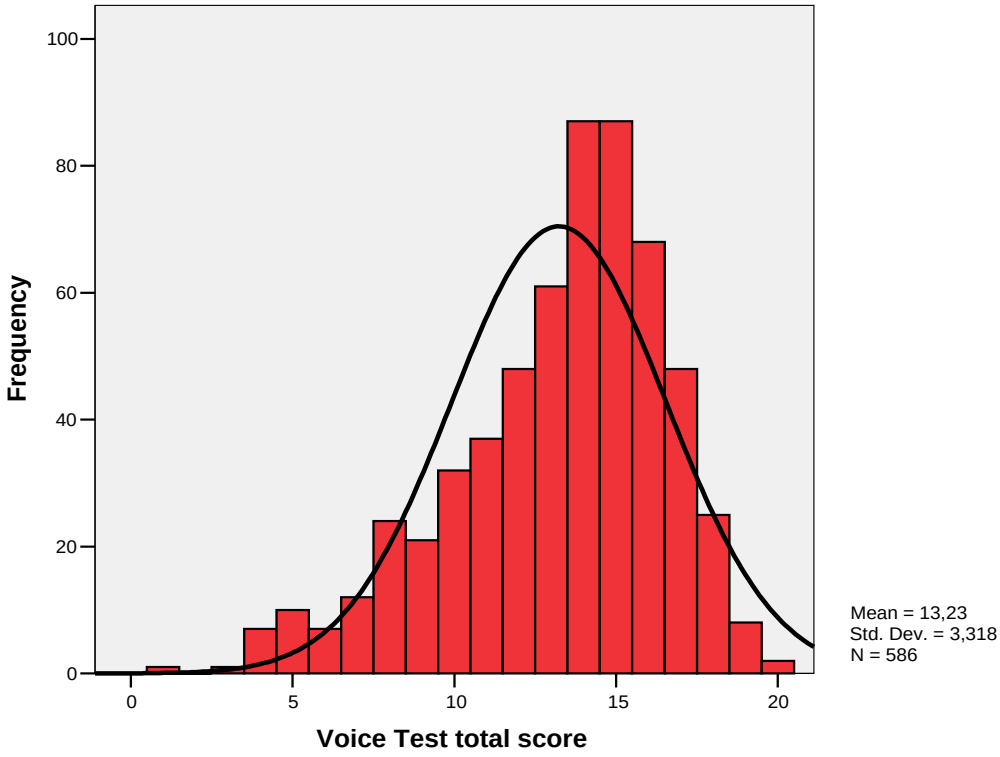


Table 2.5. Mean and standard deviation of the Voice Test total score in each group of age

	Mean	*s.d.*
78-89 months (7 years)	10,58	3,60
90-101 months (8 years)	12,12	3,03
102-113 months (9 years)	13,87	2,86
114-125 months (10 years)	14,99	1,87
126-137 months (11 years)	15,27	2,34

Influence of gender and age

Children's gender was shown to have a significant effect upon children's performance in the *Voice Test*: t (584) = 2,437, p = 0,01. Females had better performance (M = 13,57) than males (M = 12,90). The independent t test run for each group of age showed that gender difference began to become significant from 9 years old: t (134) = 1,915, p = 0,05; t (121) = 2,518, p = 0,01; t (77) = 2,745, p < 0,01, respectively for 9, 10 and 11 years old (Chart 2.9.).

In order to evaluate differences in the *Voice Test* among the different groups of age a one-way ANOVA was run with the five levels of age as the factor and the *Voice Test* total score as the dependent variable. This analysis revealed a significant effect of age on the performance in the test (Chart 12.10.), the *Voice Test* score increased with age: F (4, 581) = 55,284, p < 0,001.

Post-hoc test (Bonferroni t test) revealed significant differences among each group of age, except between 10 and 11 years old (Table 2.6.). There was not a significant interaction effect between age and gender on the Voice Test.

Chart 2.9. Gender differences in the Voice Test

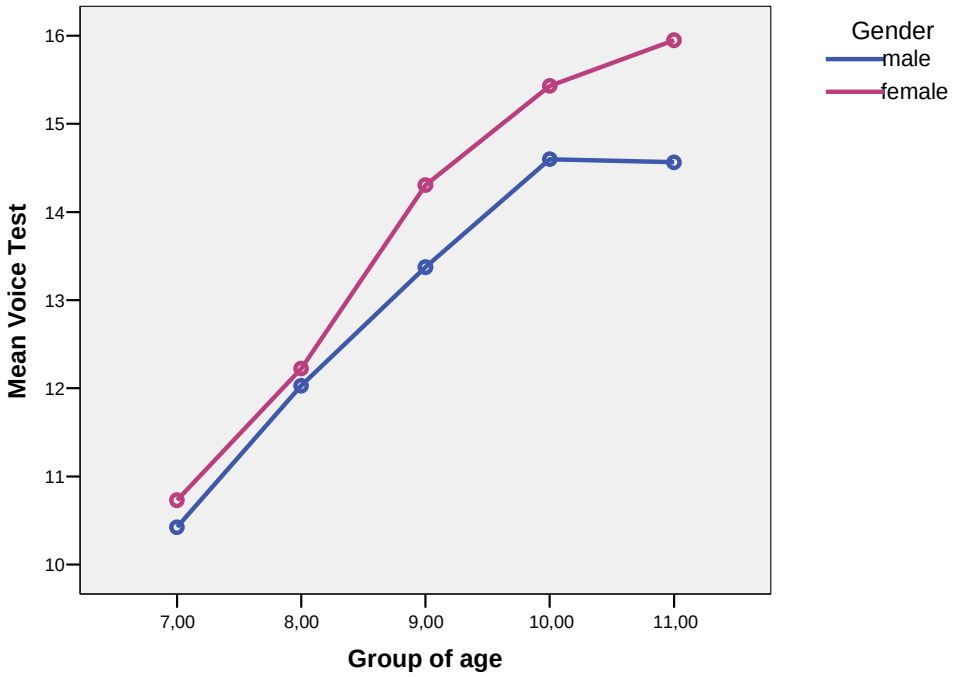


Chart 2.10. Age differences in the Voice Test

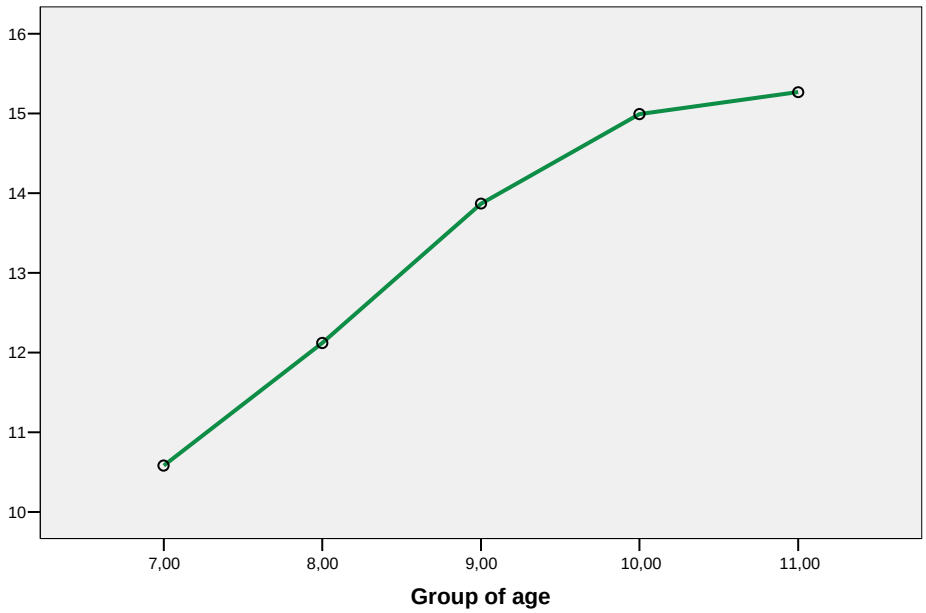


Table 2.6. Post hoc analysis

Group of age		*Mean difference (s.d.)*	*Sig.*
7 years	8 years	- 1,54 (0,36)	$p < 0,001$
	9 years	- 3,29 (0,35)	$p < 0,001$
	10 years	- 4,41 (0,36)	$p < 0,001$
	11 years	- 4,68 (0,41)	$p < 0,001$
8 years	7 years	1,54 (0,36)	$p < 0,001$
	9 years	- 1,75 (0,35)	$p < 0,001$
	10 years	- 2,87 (0,36)	$p < 0,001$
	11 years	- 3,15 (0,41)	$p < 0,001$
9 years	7 years	3,29 (0,35)	$p < 0,001$
	8 years	1,75 (0,35)	$p < 0,001$
	10 years	- 1,12 (0,35)	$p = 0,01$
	11 years	- 1,40 (0,40)	$p < 0,01$
10 years	7 years	4,41 (0,36)	$p < 0,001$
	8 years	2,87 (0,36)	$p < 0,001$
	9 years	1,12 (0,35)	$p = 0,01$
	11 years	- 0,27 (0,40)	n.s.
11 years	7 years	4,68 (0,41)	$p < 0,001$
	8 years	3,15 (0,41)	$p < 0,001$
	9 years	1,40 (0,40)	$p < 0,01$
	10 years	0,27 (0,41)	n.s.

Normative data

The *Voice Test* scores, for each group of age, were transformed in z-scores, defined by the formula: (X – M) / s.d.[5]. The magnitude pf the z-score indicates how many standard deviations is the score away from the mean value and the sign (positive or negative) of the z-score shows whether the *Voice Test* score is above or below the mean. Tables 2.7., 2.8., 2.9., 2.10. and 2.11. show the correspondent z-score for each *Voice Test* score in the five groups of age.

[5] X is the Voice Test score, M is the mean and s.d. is the standard deviation.

Table 2.7. Age 7 years (78-89 months): z-scores

Voice Test score (X)	z	*Cumulative %*
1	-2,66	0,8
3	-2,11	1,6
4	-1,83	5,7
5	-1,55	9,8
6	-1,27	12,3
7	-0,99	19,7
8	-0,72	30,3
9	-0,44	41,8
10	-0,16	47,5
11	0,12	55,7
12	0,39	67,2
13	0,67	76,2
14	0,95	84,4
15	1,23	91,8
16	1,51	97,5
17	1,78	98,4
18	2,06	100,0

Table 2.8. Age 8 years (90-101 months): z-scores

Voice Test score (X)	z	*Cumulative %*
4	-2,68	1,6
5	-2,35	5,6
6	-2,02	7,1
7	-1,69	8,7
8	-1,36	11,9
9	-1,03	15,1
10	-0,70	26,2
11	-0,37	37,3
12	-0,04	46,8
13	0,29	62,7
14	0,62	81,0
15	0,95	88,9
16	1,28	96,0
17	1,61	99,2
18	1,94	100,0

Table 2.9. Age 9 years (102-113 months): *z*-scores

Voice Test score (X)	*z*	*Cumulative %*
6	-2,75	1,5
7	-2,40	2,2
8	-2,05	5,1
9	-1,70	7,4
10	-1,35	13,2
11	-1,00	20,6
12	-0,65	29,4
13	-0,30	39,7
14	0,05	54,4
15	0,40	70,6
16	0,74	80,1
17	1,09	91,2
18	1,44	97,8
19	1,79	100,0

Table 2.10. Age 10 years (114-125 months): *z*-scores

Voice Test score (X)	*z*	*Cumulative %*
10	-2,67	1,6
11	-2,13	3,3
12	-1,60	9,8
13	-1,06	19,5
14	-0,53	36,6
15	0,01	61,0
16	0,54	81,3
17	1,07	91,1
18	1,61	97,6
19	2,14	99,2
20	2,68	100,0

Table 2.11. Age 11 years (126-137 months): z-scores

Voice Test score (X)	z	*Cumulative %*
8	-3,11	3,8
10	-2,25	5,1
11	-1,82	6,3
12	-1,40	8,9
13	-0,97	13,9
14	-0,54	30,4
15	-0,12	50,6
16	0,31	68,4
17	0,74	88,6
18	1,17	94,9
19	1,59	98,7
20	2,02	100,0

2.4.3. Discussion and conclusion

The present study provided the *Voice Test* standardized scores for Italian school age children, from 78 to 137 months. Analyses showed that the *Voice Test* score increased with age and that it can discriminate well among age, in particular among seven, eight, nine and ten years old; while ten and eleven years old children performed fairly similarly.

The study confirmed that females are better than males in Theory of Mind (Charman, Ruffman & Clements, 2002; Baron-Cohen, 2003), in fact they had higher performance than males in the *Voice Test*. This gender difference begins to become evident and significant in older children, from nine years old.

The normative data of the *Voice Test* were calculated through the transformation in z-scores for each group of age: 6,5-7,4 years; 7,5-8,4 years; 8,5-9,4 years; 9,5-10,4 years; 10,5-11,4 years.

2.5. Further analysis on the Voice Test reliability

Reliability refers to the degree to which a test is consistent and stable in measuring what it is intended to measure; it is the consistency of a measuring instrument. A test is therefore reliable if it is consistent within itself and across time. In the paragraph 2.3. I analyzed, through the Cronbach's *alpha*, a form of reliability called internal consistency reliability, that is the assessment of the consistency of results across items within the *Voice Test*.

In this paragraph I analyze further the *Voice Test* reliability, using the method of the test-retest. It consists in the administration of the *Voice Test* to the same sample in two different occasions, spaced out by a suitable period of time (not too much soon or later, because the retest could be distorted by a learning effect o by developmental change in children; Boncori, 2006). This type of reliability is very important especially for test, like the *Voice Test*, that assesses different aspects of a construct (Guilford, 1965), for example the *Voice Test* regards the comprehension of a wide range of complex mental states, both epistemic and emotional.

2.5.1. Method

Participants

Participants were 141 children (75 males and 65 females), including:

- 37 children (21 males, 16 females) aged 78-89 months (7 years old);
- 29 children (15 males, 14 females) aged 90-101 months (8 years old);
- 32 children (20 males, 12 females) aged 102-113 months (9 years old);

- 25 children (13 males, 12 females) aged 114-125 months (10 years old);
- 18 children (7 males, 11 females) aged 126-137 months (11 years old).

Children came from a middle-class background and attended the primary school in a town near Milan. Parents gave their written consent to allow children's participation in this study.

Measures

Children were administered in group the *Voice Test*, in a quiet room, following the procedure described in the previous paragraph and in the Appendix.

In addition to the Voice Test, children were tested a standardized measure of receptive vocabulary, the *Peabody Picture Vocabulary Test – Revised* (PPVT-R) (Dunn & Dunn, 1981; Italian validation: Stella, Pizzoli & Tressoldi, 2000), already described in the paragraph 2.3.1. Non verbal intelligence quotient was also assessed using the *Raven Coloured Progressive Matrix* (RCPM) (Raven, 1984).

Procedure

The *Voice Test* was administered in group of 10 children. After four weeks (considered the better period between the two test sessions, in fact children unlikely could improve they mindreading ability in few weeks and they unlikely could remember the test, reducing a lot the learning effect) children were re-tested, exactly in the same way of the previous administration. Half the sample received the PPVT – R and the RCPM individually after the *Voice Test* administration in the test condition, and half received it individually in the re-test condition.

2.5.2. Results

Descriptive statistics

Table 2.12. and 2.13. show means and standard deviations of the *Voice Test* scores, both in the test and re-test conditions, in the entire sample and in each group of age. All children recognized correctly the speakers' gender in the control test.

PPVT – R and RCPM scores were normally distributed and the had, respectively, a mean of 103,67 (s.d.: 14,35) and of 108,10 (s.d.: 14,55).

Verbal ability and IQ influences

PPVT – R correlated positively with the *Voice Test* both in test (Chart 2.11.) and re-test conditions (r = 0,431 and 0,467 respectively, $p < 0{,}001$), also controlling for age (r = 0,342 and 0,383 respectively, $p < 0{,}001$). RCPM did not correlate with the Voice Test (Chart 2.12.) in both conditions. A hierarchical regression analysis was conducted to assess the relative contribution of age, verbal ability and IQ to children's performance in the *Voice Test* (test condition). The model explained the 35% of the variance (adjusted R squared: 0,351), F (3, 137) = 25,255 ($p < 0{,}001$). The equation was: *Voice Test* score = -1,53 + 0,44 age + 0,29 verbal ability + 0,025 IQ. Only the IQ independent variable did not achieve significance, while age and verbal ability were significant ($p < 0{,}001$).

Test-retest

Pearson's product-moment correlation was calculated between the Voice Test scores obtained in the test and re-test conditions. It was 0,70 ($p < 0{,}001$).

It remained significant also controlling for verbal ability (PPVT – R) [$r = 0{,}621$, $p < 0{,}001$], for age [$r = 0{,}570$, $p < 0{,}001$] and for verbal ability and age [$r = 0{,}505$, $p < 0{,}001$].

Table 2.12. Descriptive statistics of the Voice Test in the test and re-test conditions.

	Mean	*s.d.*
Test	14,26	3,09
Re-test	15,25	3,15

Table 2.13. Descriptive statistics in the test and re-test conditions in each group of age.

Age	*Condition*	*Mean*	*s.d.*
7-year	*Test*	12,24	3,27
	Re-test	12,89	3,44
8-year	*Test*	13,52	3,17
	Re-test	14,55	2,81
9-year	*Test*	15,13	2,56
	Re-test	16,06	2,27
10-year	*Test*	15,32	1,93
	Re-test	16,60	2,16
11-year	*Test*	16,61	1,82
	Re-test	17,89	1,78

Chart 2.11. Correlation between the Voice Test and the PPVT - R

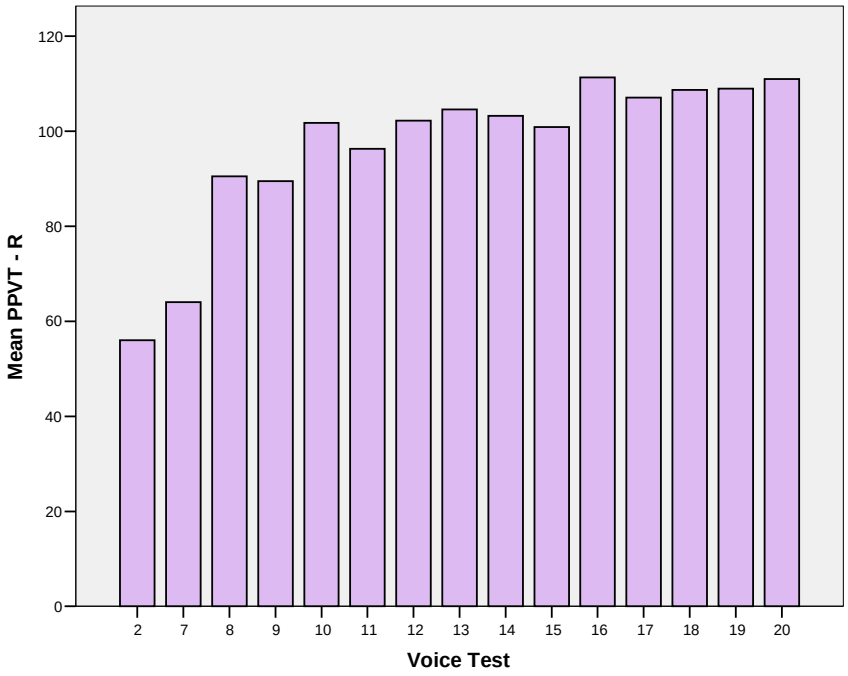
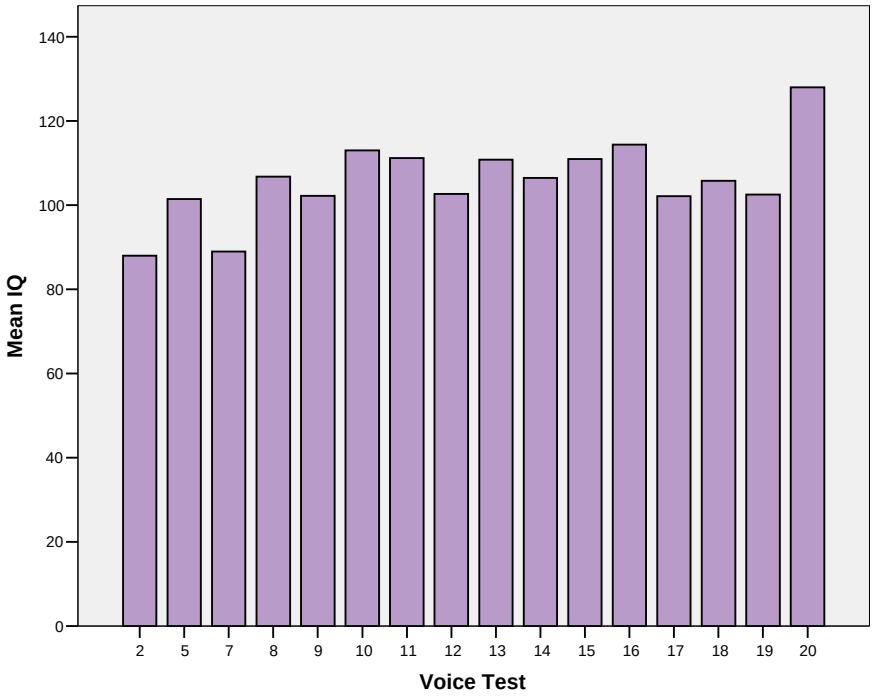


Chart 2.12. Correlation between the Voice Test and the RCPM



2.5.3. Discussion and conclusion

The present study provided some further supports to the *Voice Test* reliability, through the test-retest method. The test score was highly correlated, in the same school age sample, to the test score after 4 weeks.

Moreover the research investigated the relationship between the *Voice Test* and two standardized measures of verbal (*PPVT – R*) and non-verbal (*RCPM*) ability. As in the previous study it was found that the *Voice Test* is strictly linked to the receptive language ability (paragraph 2.3.); the novel finding was that the non-verbal ability was not related to the performance in the *Voice Test*.

These data suggest that Theory of Mind is independent from the non-verbal intelligence quotient, while it is interdependent in development with verbal ability (e.g., Antonietti, Liverta Sempio & Marchetti, 2006; Milligan, Astington & Dack, 2007). The regression analysis confirmed this hypothesis, in fact only age and verbal ability were significant predictors of the *Voice Test* performance.

In conclusion, this research contributed to further validate the new Theory of Mind advanced instrument based on auditory cues created, the Italian *Voice Test* – child version, through the analysis of the test-retest reliability and the relationships with verbal and non verbal ability.

2.6. Conclusion: Is the Voice Test a suitable ToM measure for school age children?

The current research describes the three phases of the development, validation and standardization of a new Italian Theory of Mind advanced task, based of perceptive cues, for school age children (78-137 months): the *Voice Test*. The test, that can be administered in a group setting (ten children), consists of 20 auditory stimuli, that are spoken phrases with a neutral meaning. Children were asked to understand speakers' complex mental states, choosing it among four mental words. In the control trial they were asked to recognize speakers' gender. To summarize, the results indicate that the psychometric qualities (internal and test-retest reliability; construct validity) of the test are good and provide the normative data for a normal children population aged from 6,5 to 11,4 years.

In particular, the internal consistency is acceptable: Cronbach's *alpha* is 0,69. This value can be influenced by the heterogeneity of mental states assessed, in fact the test regards both epistemic and complex emotional mental states. The test-retest reliability was high. The construct validity of the test was demonstrated through the positive correlations between the *Voice Test* and others classical (second-order *False Belief* task) and advanced (*Eyes Test*) ToM tasks. These results suggested that the construct of ToM is at the same time homogeneous, because of the strong correlations, and wide, because it comprises various types of mental states (i.e., epistemic states in the false belief task, epistemic and emotional states in the *Eyes Test*, epistemic and complex emotional states in the *Voice Test*) and different mental processes (i.e., different modalities used to recognize mental states, that use the cognitive, the visual or the auditory channel).

The *Voice Test* performance increase with age: significant differences were found among 7, 8, 9 and 10, while children aged 11 have a performance similar to those aged 10. Females obtained higher score in the test than males, as was found in other ToM tasks (Charman, Ruffman & Clements, 2002; Baron-Cohen, 2003).

The correlation between the *Voice Test* and standardized IQ measures showed that the test correlated with and was predicted by the receptive vocabulary ability (assessed with the PPVT – R), confirming results obtained with other classical and advanced ToM tests and the link between ToM and language ability (Astington & Baird, 2005; Antonietti, Liverta Sempio & Marchetti, 2006; Milligan, Astington & Dack, 2007).

Non verbal IQ, assessed with RCPM, did not correlate with the *Voice Test* showing its independence from intellectual ability, as found in other advanced ToM tasks based on perceptive cues (e.g., *Eyes Test* by Baron-Cohen et al., 1997, 2001a; *Voice Test* by Rutherford, Baron-Cohen and Wheelwrigth, 2001; *Reading the Mind in the Films* by Golan et al., 2006b; see paragraph 2.1.)

In conclusion the *Voice Test* can be considered a valid ToM measure, suitable for school age children. It will can be used to assess ToM in older children, besides other ToM tasks, in order to better grasp the complexity of mindreading ability during development. It will can be useful to study ToM in children with disabilities, such as children with autism or language impairment, but also blind children in the auditory communicative channel is very important because they cannot count on the visual one.

The *Voice Test* can also contribute to analyze individual differences in ToM, showing how and why different ToM abilities develop; for example mental state understanding from eyes, from voices and from the "cognitive" channel (e.g., knowledge of the situation) can differ both among children and within the children themselves.