Cognitive Development 37 (2016) 28–41

Contents lists available at ScienceDirect

Cognitive Development

Continued development of recursive thinking in adolescence: Longitudinal analyses with a revised recursive thinking test

Esther van den Bos[a,*], Mark de Rooij[a], Sindy R. Sumter[b], P. Michiel Westenberg[a]

[a] *Institute of Psychology, Leiden University, The Netherlands*
[b] *Amsterdam School of Communication Research, University of Amsterdam, The Netherlands*

ARTICLE INFO

Article history:
Received 14 August 2014
Received in revised form 21 October 2015
Accepted 20 November 2015

Keywords:
Social cognition
Adolescence
Longitudinal study
Recursive thinking
Verbal intelligence
Psychometric properties

ABSTRACT

The present study adds to the emerging literature on the development of social cognition in adolescence by investigating the development of recursive thinking (i.e., thinking about thinking). Previous studies have indicated that the development of recursive thinking is not completed during childhood. The present study focused on late childhood and adolescence and presents the first longitudinal data on recursive thinking. At Time 1, 299 participants, aged 8 to 17 years, completed a revised version of the recursive thinking test developed by Miller, Kessel and Flavell (1970). At Time 2, two years later, 221 participants completed the test again. Psychometric properties of the revised test were found to be adequate. The developmental analysis showed that scores increased with age—both between- and within participants-, indicating that recursive thinking continues to develop throughout adolescence and does not level off before 18 years of age. Verbal abilities only partially explained this development.

1. Introduction

Interest in the development of social cognition during adolescence has been growing since neuroscience studies have demonstrated that brain regions associated with social cognition continue to develop throughout adolescence (Blakemore & Choudhury, 2006; Burnett & Blakemore, 2009; Nelson, Leibenluft, McClure, & Pine, 2005). These findings have raised the question how the changes affect social cognition and generated the hypothesis that some socio-cognitive processes should continue to develop during adolescence. The present study investigated whether recursive thinking is such a process.

Recursive thinking, or thinking about thinking (Miller, Kessel, & Flavell, 1970), is required to infer other people's thoughts. It is an aspect of social cognition that has been related to (cognitive) perspective taking (Landry & Lyons-Ruth, 1980; Veith, 1980) and theory of mind (Miller, 2012). For example, second-order (false) beliefs can be described as recursive thinking: John thinks that Mary thinks that the ice-cream van is in the park (Perner & Wimmer, 1985). Recursive thinking can be considered a specific form of Theory of Mind (ToM), restricted to epistemic states (Perner & Wimmer, 1985), whereas the broader ToM concept encompasses thinking about other mental states, such as desires, intentions (e.g., Blakemore & Choudhury, 2006) and feelings (e.g., Baron-Cohen, Wheelwright, Hill, Raste, & Plumb, 2001).

* Corresponding author at: Leiden University, Developmental Psychology Unit, PO Box 9555, 2300 RB Leiden, The Netherlands. Fax: +31 71 527 3619.
E-mail address: bosejvanden@fsw.leidenuniv.nl (E. van den Bos).

http://dx.doi.org/10.1016/j.cogdev.2015.11.002
0885-2014/

The ability to think about thinking has been suggested to underlie successful social performance in domains such as the display of emotions (Saarni, 1979), self-presentation (Miller, 2012), persuasion (Hill & Palmquist, 1978), moral reasoning (Hill & Palmquist, 1978S; Miller, 2012; Veith, 1980) and communication (Miller, 2012). For example, Perner and Wimmer (1985) noted that children (under 8 years of age) may answer test questions incompletely, because they do not understand that the examiner wants to know what they know. In line with these proposals, several empirical studies have related recursive thinking to effective social functioning. Stiller and Dunbar (2007) demonstrated that adults who are better at recursive thinking maintain larger social networks, naming more persons they could turn to for support in case of personal problems. High performance on recursive thinking tests was also positively related to agreeableness (Nettle & Liddle, 2008; Ferguson & Austin, 2010), emotional intelligence as a trait and in specific situations (Ferguson & Austin, 2010) and self-reported cooperativeness (Paal & Bereczkei, 2012), though not to successful recognition of cooperativeness in others (Sylwester, Lyons, Buchanan, Nettle, & Roberts, 2012). Low performance on recursive thinking tests was associated with a tendency to blame other people for negative events (Kinderman, Dunbar, & Bentall, 1998) and to interpret their intentions as hostile (Jeon et al., 2013).

Recursive thinking has long been assumed to emerge in childhood and to become more advanced in adolescence. *Indirect* support for the assumption comes from two studies on the development of related abilities. First, Miller (2012) reviewed findings from higher-order ToM tasks in which participants had to answer questions about a story character's beliefs about other story characters' beliefs (e.g., Kinderman et al., 1998). He concluded that, from childhood to adulthood, participants become able to deal with increasingly longer chains of beliefs. Second, recursive analogical reasoning, which is solving analogies between analogies (e.g., sand stands to beach as star stands to galaxy is analogous with water stands to ocean as air stands to sky), improved from grade five to seven and from grade seven to nine (ages 10–15), though not from grade nine to eleven (ages 14–17; Nippold, 1994). However, there is little *direct* evidence that recursive thinking continues to develop throughout adolescence. Several studies have demonstrated the emergence of recursive thinking in childhood (Eliot, Lovell, Dayton, & McGrady, 1979; Landry & Lyons-Ruth, 1980; Miller et al., 1970; Oppenheimer, 1986; Veith, 1980), but only one study included adolescents (Müller & Overton, 2010). It demonstrated further improvement in early adolescence, but not beyond.

1.1. Development of recursive thinking

The study of recursive thinking was initiated by Miller et al. (1970). They developed a test of children's understanding that representational actions, unlike physical actions, can be self-embedded (e.g., *I think that she thinks that I think that* her story is not interesting). Specifically, they hypothesized that the development of this understanding would follow an invariant sequence of (a) thinking about contiguous people (i.e., thinking about one or more persons), (b) thinking about an action between people (e.g., thinking about someone talking to another person), (c) one-loop recursive thinking (i.e., thinking about someone who is thinking about someone) and (d) two-loop recursive thinking (i.e., thinking about someone who is thinking about someone who is thinking about someone). They tested the hypothesis using a cartoon description task. Participants were presented with cartoons that contained thought clouds, speech bubbles and up to four different characters (a boy, a girl, a father and a mother). In each cartoon, the boy was depicted with a thought cloud over his head, in which all other elements were embedded. The participant had to tell what the boy was thinking. (See Appendix A for a two-loop recursion item used in the present study).

Miller et al. (1970) tested elementary school children from grades one to six (i.e., ages 6–12). They found that, in all grades, accuracy was higher for contiguity items than for action items, higher for action items than for one-loop recursion items and higher for one-loop recursion items than for two-loop recursion items. This pattern held for group means as well as the individual scores of 91.7% of the participants. Miller et al. (1970) concluded that the four types of items were scalable (i.e., constituting a scale with four distinct levels of difficulty), representing consecutive steps in the development of the understanding of recursive thinking. They also charted this development from grade one to six (i.e., ages 6–12). Performance on contiguity items was already at ceiling in first grade, but performance on the other items increased with grade level (percentages were derived from Fig. 2 in Miller et al., 1970): (a) accuracy on action items increased from about 43% in first grade to about 85% in fourth grade and remained fairly stable thereafter; (b) accuracy on one-loop recursion items increased from about 19% in first and second grade to about 46% in third and fourth grade and about 57% in fifth and sixth grade; (c) accuracy on two-loop recursion items was below 10% in grades one through three, about 17% in grade four and about 36% in grades five and six. Subsequent studies administering the cartoon description task replicated the developmental pattern in six to ten-year-olds (Eliot et al., 1979; Landry & Lyons-Ruth, 1980; Oppenheimer, 1986; Veith, 1980). The results indicate that the development of recursive thinking is still in progress in late childhood, in particular for one-loop and two-loop recursion.

Müller and Overton (2010) extended the study of recursive thinking into adolescence. In a study with participants from grades two, four, six and eight (ages 6–14), they largely replicated the results for grades two to six reported by Miller et al. (1970). Moreover, they demonstrated that performance on two-loop items continued to improve from grade six to grade eight (ages 11–14). In another study with participants from grades five, eight and eleven (ages 10–18) they found no further improvement between grades eight and eleven (ages 13–18) on action, one-loop recursion and two-loop recursion items. However, considerable variability in performance of the eight graders across the two studies prevents the conclusion that the development of recursive thinking levels off in early adolescence. A more fine-grained examination is needed to clarify the pattern of development.

1.2. Levels of recursive thinking and the relation with verbal abilities

Müller and Overton (2010) also tested the claim that the levels of recursive thinking proposed by Miller et al. (1970) are qualitatively different. Using Rasch analysis, they demonstrated that the cartoon description task measured a single dimension with three distinct levels: thinking about contiguous people, thinking about action and recursive thinking. One-loop and two-loop recursive thinking were not identified as separate levels. While supporting qualitative differences between the first three levels, this finding raises the possibility that one-loop and two-loop recursive thinking are not different in kind.

Another unresolved issue concerns the relation with individual differences in intellectual capacity. As noted already by Miller et al. (1970), the cartoon description task is a verbal task and performance is likely to be related to one's verbal intelligence. Yet, very limited information is available. Children's score on the recursive thinking test has been found to correlate moderately with IQ (Veith, 1980). Müller and Overton (2010) examined whether linguistic complexity could explain the observed order of difficulty of their items and concluded that this was not the case. As discussed above, they found that action items were easier than one-loop and two-loop recursion items, with no difference between the latter two types of items. Linguistic complexity, in contrast, was similar for action items and one-loop recursion items, but higher for two-loop recursion items.

1.3. The present study

In summary, the development of recursive thinking has been proposed to continue during adolescence (e.g., Miller et al., 1970). In line with this suggestion, it has been demonstrated that recursive thinking is not fully developed by the end of childhood. However, only limited support has been provided for development in early adolescence and there is no support for further development. The current study specifically focused on late childhood and adolescence and presents the first longitudinal data on the development of recursive thinking. The primary aim was to investigate whether the development of recursive thinking levels off in early adolescence, as suggested by the results of Müller and Overton (2010), or continues throughout adolescence, as may be expected from the development of brain areas associated with social cognition (Blakemore & Choudhury, 2006; Burnett & Blakemore, 2009; Nelson et al., 2005). The cohort-sequential design of our study allowed us to directly address this issue by testing whether participants' improvement over time depended on their age. In addition, our longitudinal data allowed for an exploration of the role of verbal abilities in the development of recursive thinking. Before investigating these two issues, we first examined the psychometric properties (factor structure, scalability, reliability, discriminant validity with verbal intelligence) of the revised recursive thinking test, which we had adapted for use with an older population. Examination of the factor structure also provided evidence on the levels of recursive thinking that can be distinguished.

2. Method

2.1. Participants

The data used in the present study are part of the Social Anxiety and Normal Development (SAND) study. The aims of the SAND-study were to chart the development of social anxiety as well as normative developments in the physical and socio-cognitive domains in a community sample of adolescents and to investigate how social anxiety and stress responses to public speaking are related to individual difference variables (see e.g., Miers, Blöte, De Rooij, Bokhorst, & Westenberg, 2013) as well as normative developments in the physical (see e.g., Van den Bos, De Rooij, Miers, Bokhorst, & Westenberg, 2014) and socio-cognitive domains (Van den Bos, Van Duijvenvoorde, & Westenberg, submitted for publication). The SAND-study was approved by the Leiden University Medical Ethical Committee and carried out in accordance with the Declaration of Helsinki. Parents provided active consent; written assent was obtained from participants themselves.

Participants were recruited through two elementary schools and one secondary school in Leiden, a middle-sized city in the Netherlands. The majority of participants were of Dutch origin: 93.5% of the participants and 87.4% of their parents was born in the Netherlands. The sample included 126 elementary school children and 173 adolescents from all educational streams in the Dutch school system representing varied levels of intelligence in the whole sample and within all age groups. For more information on the sample see Van den Bos et al. (2014).

The data for the SAND-study were collected in four waves, with the first wave being fielded in 2006. The recursive thinking test was administered in Wave 1 and Wave 3. For clarity, these data collection points will be referred to as Time 1 and Time 2 in the present study. At Time 1, there were 299 participants: 154 males (51.5%) and 145 females (48.5%). Their ages ranged from 8 to 17 years ($M = 13.2$, $SD = 2.3$). At Time 2, two years later, 222 participants returned (51.4% male, mean age = 15.0, SD = 2.2). The attrition rate was 25.8%. There was no difference in the distribution of gender ($\chi^2(1) = .008$, $p = .928$) or mean age at Time 1 ($t(297) < 1$) between those who continued to participate at Time 2 and those who did not. The mean score on the test of recursive thinking was higher for participants who returned at Time 2 than for participants who did not return ($t(297) = 2.518$, $p = .012$). However, an independent samples median test showed that the median scores were not significantly different for the two groups ($p = .125$), indicating that the lower mean for participants who did not return was due to a small number of extremely low scores.

2.2. Overall procedure

At both Time 1 and Time 2, there were two sessions of data collection in the lab: a pre-session and a public speaking session (see Westenberg et al., 2009). The test of recursive thinking was administered during the pre-session, together with various self-report questionnaires, a sentence completion test measuring psychosocial development and three other cognitive tests: the Six Part Test (Emslie, Wilson, Burden, Nimmo-Smith, & Wilson, 2003) and the Vocabulary and Analogies subtests of the Differential Aptitude Test (DAT; De Wit & Compaan, 2005). The DAT Vocabulary was only administered during the pre-session at Time 2. At Time 1, it was administered during an additional session of data collection in a classroom of the participant's school.

2.3. Instruments

2.3.1. Recursive thinking test

An adapted version of the cartoon description task developed by Miller et al. (1970) was used. Contiguity items were omitted in this version, because ceiling effects have been found with children from second grade elementary school and higher (Miller et al., 1970; Müller & Overton, 2010; Oppenheimer, 1986). Participants were presented with 14 cartoons: 3 action items, 5 one-loop recursion items and 6 two-loop recursion items. The items were those used by Miller et al. (1970) and two additional two-loop recursion items (see Appendix A). The materials were newly created professional drawings (see Fig. A1 for an example), printed on A4-sized laminated paper. Clouds with smooth outlines represented talking; clouds with scalloped outlines represented thinking. The characters in the cartoons were a boy, a girl, a man and a women. Persons talking to each other in the action items were depicted looking at each other, in response to a recommendation by Miller et al. (1970). The full set of cartoons is available from the authors upon request.

2.3.1.1. Procedure. Each participant was tested individually and an (MP3) audio recording was made during the whole session. The experimenter first presented an example of a speech cloud and a thought cloud and checked whether the participant understood the difference. Next, the experimenter presented a drawing of the four characters and labeled them as "the boy", "the girl", "the male teacher" ("meester" in Dutch) and "the female teacher" ("juf" in Dutch). The man and the woman were referred to as teachers to fit the overall context of the SAND-study. It was stressed that only four individuals could appear in the cartoons (i.e., there were no twins). The participant had to name the four characters on a drawing without labels. During the actual test, male participants were presented with cartoons in which the boy was the main character and female participants were presented with cartoons in which the girl was the main character. The cartoons were presented one at a time, in the following order: 1, 3, 4, 5, 2, 6, 10, 8, 11, 14, 12, 13, 9, 7. For each cartoon, the participant had to answer the question "What is the boy (girl) thinking?" If the participant gave an ambiguous answer, the experimenter asked an open question for clarification and prompted the participant to repeat the description. The prompts were standardized. The best of the two descriptions was scored (no further attempts were allowed for that item).

2.3.1.2. Scoring. The audio recordings were transcribed and scored by two independent raters. In contrast to the binary (0 = incorrect, 1 = correct) scoring procedure described by Miller et al. (1970), scores of 0, 1 or 2 were assigned in the present study. A score of 0 was assigned when the description did not reflect the right level of recursive thinking (e.g., not including a cognitive verb, not recursive, too few or too many loops of recursion). A score of 1 was assigned when the description reflected the right level of recursive thinking, but was not entirely accurate. This score was mainly assigned to descriptions in which the characters were labeled ambiguously (e.g., "the boy is thinking about the boy/him" instead of "himself", using a pronoun with unclear reference). A score of 2 was assigned to correct descriptions. A detailed scoring manual (in Dutch) is available upon request. The trinominal scoring system was introduced to distinguish between ambiguous answers that may have been generated according to a syntactic strategy without mentally representing the situation (scored as 1) and answers reflecting a full mental representation of the situation (scored as 2). The latter two scores would be rated as correct (=1) in the binary system. The total and subscale scores obtained with the trinomial system are highly correlated with the total and subscale scores obtained with the binary scoring system (collapsing scores 1 and 2). At Time 1 total score $r = .95$, action score $r = .91$, one-loop recursion score $r = .95$, two-loop recursion score $r = .96$. At Time 2 total score $r = .94$, action score $r = .79$, one-loop recursion score $r = .93$, two-loop recursion score $r = .96$. All correlations were significant at $\alpha = .001$. All statistical analyses in this paper were done on the trinominal scores. To facilitate comparison with previous studies, proportions correct according to the binominal system are provided in Supplementary Table 1 for each age group.

Agreement between the two raters was good. Cohen's Kappa ranged from .66 to .93 (median = .82) across the 14 items at Time 1 and from .51 to .90 (median = .82) at Time 2. Kappa was lowest for the action items. This can be explained by the high prevalence of score 2 (and low prevalence of scores 0 and 1) on these items. Kappa is corrected for agreement by chance. When the scores have a very different prevalence, chance agreement is high and kappa is reduced (Sim & Wright, 2005). The Kappa of .51 reflects disagreement on 14 out of 221 cases with a prevalence of 17 for score 0, 13 for score 1 and 412 for score 2. In case of disagreement, the raters discussed the description and reached a consensus score. Scores on the recursive thinking test were available for all 299 participants at Time 1 and for 221 of the 222 participants at Time 2. The test could not be administered to one participant, because she was unable to come to the lab.

2.3.2. Differential aptitude test (DAT): vocabulary and analogies

Two subtests of the Dutch version of the Differential Aptitude Test (De Wit & Compaan, 2005) were administered to measure verbal abilities: Vocabulary and Analogies. The tests are suitable for adolescents of twelve years and older. Therefore, they were not administered to participants in primary school. The tests were presented in a booklet and participants had to circle the correct response. The DAT was designed for use in the context of vocational advice and it is normed according to grade and educational stream. Because a possible relation between verbal intelligence and recursive thinking would be unlikely to depend on educational stream, raw scores (number of correct answers) were used rather than normed scores.

2.3.2.1. DAT vocabulary. The vocabulary test measures one's knowledge of word meaning. The test comprised 60 items and scores ranged from 0 to 60. Each item consisted of a target word and five alternatives. Participants had to select a synonym of the target word from the alternatives. They were allowed to work on the vocabulary test for 15 min. The test has good internal consistency. Kuder–Richardson's (formula 20) reliability ranged from .71 to .82 across grades and educational streams in the manual (De Wit & Compaan, 2005). In the present study, Cronbach's alpha was .86 at Time 1 and .90 at Time 2. Scores were available for 173 participants at Time 1 and 189 participants at Time 2.

2.3.2.2. DAT analogies. The analogy test measures one's ability to recognize relations between words, tapping into verbal abstraction and reasoning. The test comprised 40 items and scores ranged from 0 to 40. Each item consisted of two incomplete word pairs (the first word of the first pair and the last word of the last pair were replaced by a question mark) and five alternative sets of a first word and a last word. Participants had to select the set that would create an analogy between the word pairs. Participants were allowed to work on the analogy test for 20 min. The test has good internal consistency. In the manual (De Wit & Compaan, 2005) Kuder–Richardson's (formula 20) reliability ranged from .79 to .87 across grades and educational streams. In the present study, Cronbach's alpha was .91 at Time 1 and .93 at Time 2. Scores were available for 170 participants at Time 1 and 188 participants at Time 2.

2.4. Analyses

2.4.1. Psychometric properties

In line with the hypothesis of Miller et al. (1970) and the work of Müller and Overton (2010), we hypothesized that the 14 items of the adapted version of the recursive thinking test would load on a single factor (recursive thinking), while clustering at three levels of difficulty (corresponding to thinking about action, thinking about thinking and thinking about thinking about thinking). We examined the hypothesized factor structure and scalability of the items using confirmatory factor analysis with Mplus version 6.11. It was specified that all items were ordinal. The model fit was evaluated using the Comparative Fit Index (CFI) and the Root Mean Square Error of Approximation (RMSEA). For the CFI, values between .90 and .95 are considered acceptable and >.95 as good. For RMSEA, acceptable models have values <.10 and good models have values <.05. Finally, following Yu and Muthén (2002) the Weighted Root Means Square Residual (WRMSR) was provided as an additional fit index for categorical data. A good fit is evidenced by the WRMSR being smaller than 1. To test the scalability hypothesis, the thresholds in probits for transitioning from score 0 to score 1 and from score 1 to score 2 were compared for the three types of items. Thresholds were expected to be significantly lower for action items than for other items and significantly lower for one-loop items than for two-loop items.

Reliability of the recursive thinking test was investigated by computing Cronbach's alpha. Discriminant validity was investigated by examining Pearson correlations of the mean scores on the recursive thinking test with the available sum scores on the DAT Vocabulary and the DAT Analogies. These correlations may be inflated, because both the scores on the recursive thinking test and the raw scores on the DAT are related to age. Therefore, partial correlations controlling for age were also reported to assess discriminant validity in participants of the same age.

2.4.2. Development of recursive thinking

The development of recursive thinking in adolescence was investigated with a cumulative logit model with proportional odds property (Agresti, 2007). This is a logistic regression model, which makes use of the ordering of the response categories. In the proportional odds model, a regression weight describes the effect of a variable on the log odds of response above a certain category (Agresti, 2007). In the present study, the response categories were 0, 1 or 2. Model 1 examined the effects of gender and age on the log odds of scoring above a category (i.e., above a score of 0, above a score of 1) and whether these effects varied by type of item. Items were coded as action, one-loop recursion or two-loop recursion and were not considered individually. Age was decomposed into an initial level and a change component to examine both between- and within-participants effects of age (De Rooij, 2013; Hedeker & Gibbons, 2006). The explanatory variables were type of item, gender (female = 0, male = 1), age at Time 1 centered at the age of the youngest participant (T1Age; i.e., the between subjects effect of age), the difference between the age at the time of testing and the age at Time 1 (ΔAge; i.e., the within-subjects

Table 1
Confirmatory factor analysis estimates for Time 1 and Time 2.

Item	Time 1	Time 2
1	0.52	−0.02
2	0.78	0.51
3	0.61	0.46
4	0.60	0.61
5	0.69	0.59
6	0.76	0.68
7	0.70	0.73
8	0.52	0.57
9	0.68	0.78
10	0.45	0.75
11	0.63	0.68
12	0.49	0.57
13	0.70	0.78
14	0.57	0.79

effect of age), two-way interactions between T1Age and ΔAge (i.e., non-linear effect of age), type of item and gender, type of item and T1Age, type of item and ΔAge, and the three-way interaction between type of item, T1Age and ΔAge.

Two additional analyses were performed to investigate whether the development of recursive thinking was influenced by verbal abilities. First, we tested the model described above on the data from 130 participants who were 12 years and older at Time 1 and had completed the subtests of the DAT (model 2). Then we controlled for verbal abilities by adding the scores on DAT Vocabulary and DAT Analogies and their interactions with the type of item as explanatory variables (model 3). DAT scores were included as time-varying variables, so that recursive thinking at Time 1 was controlled for verbal abilities at Time 1 and recursive thinking at Time 2 was controlled for verbal abilities at Time 2.

For statistical inference, the clustered bootstrap procedure (De Rooij, 2013; Sherman & Le Cessie, 1997) was used to combine data from the same participants at Time 1 and Time 2 in a single analysis. Intercepts and regression weights were estimated in the same way as in the standard proportional odds model, but the standard errors were derived by bootstrapping. From the total data set, 2000 bootstrap samples of the same size as the original set were drawn randomly with replacement. To deal with the dependency between measurements of the same individual, the bootstrap was clustered: individuals rather than cases were sampled, so that if an individual was assessed at both times, both measurements were included in the bootstrap sample (De Rooij, 2013). The analyses were run in R 3.0.0 (R Core Team, 2013).

3. Results

3.1. Psychometric properties

3.1.1. Factor structure

At Time 1 ($n = 299$), the fit for the single factor model was close to acceptable: $\chi^2(77) = 310.03$, $p < .001$; CFI: .85; RMSEA = .10, 95% CI: [.09–.11], and the WRMSR = 1.52. At Time 2 ($n = 221$), the single factor model had an acceptable fit: $\chi^2(77) = 172.74$, $p < .001$; CFI: .92; RMSEA = .08, 95% CI: [.06–.09], and the WRMSR = 1.14. However, one item showed negative and low loading at Time 2. The standardized estimates of the 14 factor loadings for each time point are presented in Table 1.

3.1.2. Scalability

Fig. 1 shows thresholds from 0 to 1 and from 1 to 2 with 95% confidence intervals for the items at Time 1 and Time 2. The threshold from 1 to 2 is missing for item 9, because participants' scores on this item fell into only two categories. The results were highly similar for Time 1 and Time 2. As expected, thresholds were generally lower for action items than for one-loop recursion items and lower for one-loop recursion items than for two-loop recursion items. However, there was some overlap in confidence intervals between action items and one-loop recursion items. For action item 3, the threshold from 1 to 2 was significantly higher than for the other action items and within the 95% CI of one-loop recursion items 5, 6 and 7. Conversely, for one-loop recursion item 4, the threshold from 1 to 2 was significantly lower than for the other one-loop recursion items and within the 95% CI of action items 1 and 2. For one-loop recursion item 6, the threshold from 0 to 1 was significantly lower than for the other one-loop recursion items (except for item 4 at Time 2) and within the 95% CI of the action items. The thresholds of the two-loop recursion items were all significantly higher than the thresholds of the action items and the one-loop recursion items. In conclusion, the present data provide partial support for the scalability hypothesis.

3.1.3. Reliability and discriminant validity

The recursive thinking test showed good reliability, at both Time 1 ($\alpha = .80$, $n = 299$) and Time 2 ($\alpha = .80$, $n = 221$). For participants of 12 years and older, correlations of the mean score on the recursive thinking test with the sum scores on DAT Vocabulary and DAT Analogies were low at Time 1 and moderate at Time 2 (see Table 2). Partial correlations controlling for

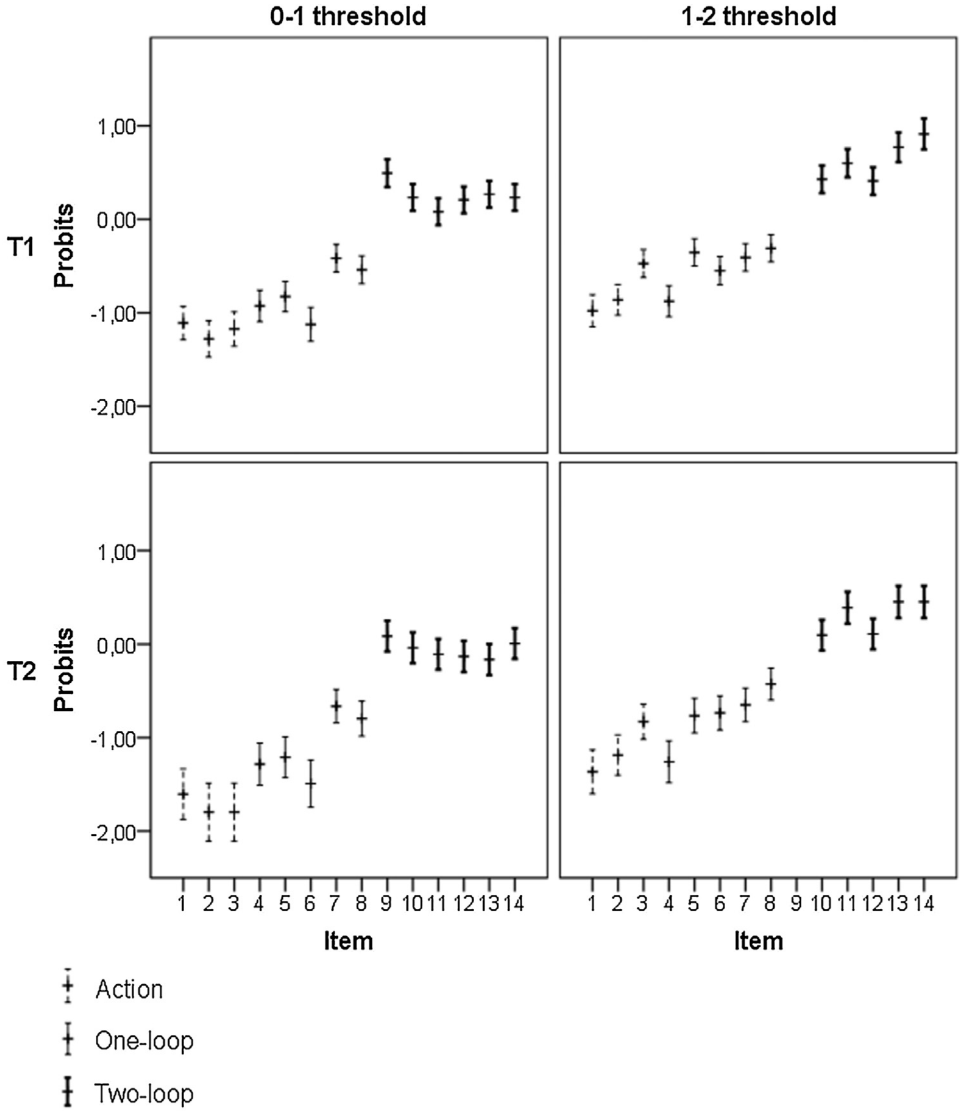

Fig. 1. Thresholds with 95% confidence intervals for transition from score 0 to score 1 and from score 1 to score 2 for the items of the recursive thinking test at Time 1 and Time 2.

age were low at both times, indicating that the recursive thinking test has good discriminant validity with verbal abilities when used with participants from a narrow age-range. The correlations between Time 1 and Time 2 scores indicated that individual differences in recursive thinking were moderately stable, whereas individual differences in verbal abilities were highly stable.

3.2. Development of recursive thinking in adolescence

Fig. 2 illustrates the increase with age in the mean scores for each type of item and for the total recursive thinking test. Mean scores were computed over items scored as 0, 1 or 2 and could range from 0 to 2. The lines represent the changes in mean scores from Time 1 (left end of line) to Time 2 (right end of line) for participants who were 9, 10, 11, 12, 13, 14, 15 and 16 years old at Time 1. Because of the small number of 17-year-olds at Time 1 ($n = 3$), these data were not plotted.

In the first analysis, scores on the 14 items were included for all 299 participants at Time 1 and all 221 participants at Time 2. Table 3 (Model 1) shows the 95% confidence intervals of the explanatory variables in the proportional odds model. An explanatory variable is significant when the confidence interval does not include zero. The main effect of type of item indicated that action items were easier than one-loop and two-loop recursion items: the log odds of scoring above each level (i.e., above a score of 0, above a score of 1) were significantly lower for one-loop and two-loop recursion items than for action items. The difference in log odds between action items and two-loop recursion items was smaller in males than in

Table 2
Correlations between the mean score on the recursive thinking test and the sum scores on the DAT vocabulary and the DAT analogies at Time 1 and Time 2.

		1	2	3	4	5
1. Recursive thinking T1	*r*	–				
	pr					
	n					
2. DAT vocabulary T1	*r*	.24**	–			
	pr	.10				
	n	173				
3. DAT analogies T1	*r*	.33***	.46***	–		
	pr	.29***	.44***			
	n	170	170			
4. Recursive thinking T2	*r*	.50***	.30**	.39***	–	
	pr	.44***	.13	.35***		
	n	221	132	130		
5. DAT vocabulary T2	*r*	.40***	.82***	.52***	.44***	–
	pr	.29***	.74***	.51***	.31***	
	n	189	132	130	189	
6. DAT analogies T2	*r*	.46***	.55***	.73***	.41***	.69***
	pr	.38***	.38***	.75***	.29***	.56***
	n	188	132	130	188	187

Note: T1 = Time 1 score, T2 = Time 2 score. Partial correlations are controlling for age.
** $p < .01$.
*** $p < .001$.

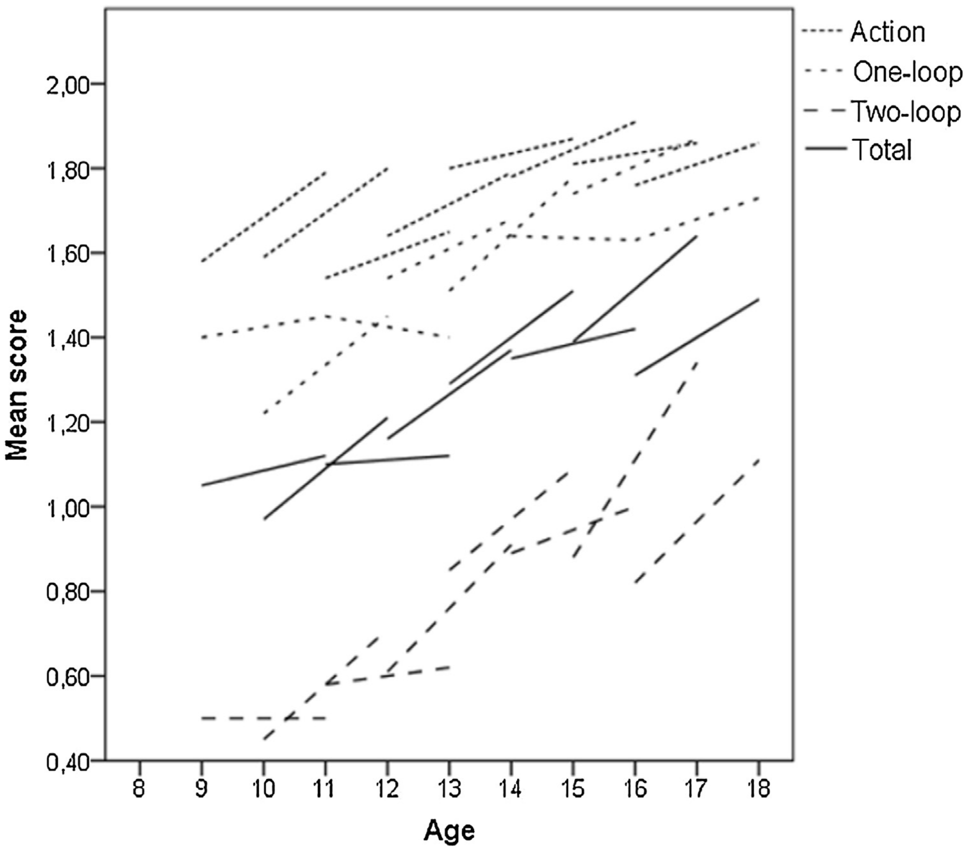

Fig. 2. Change from Time 1 (left end of each line) to Time 2 (right end of each line) in the mean scores for each type of item and the total recursive thinking test for participants who were 9–16 years old at Time 1. Mean scores were computed over items scored as 0, 1 or 2 and could range from 0 to 2.

females, as shown by the significant interaction between type of item and gender. Moreover, there were significant effects of age: the log odds of scoring above each level increased both with age at Time 1 (between-participants effect) and from Time 1 to Time 2 (within-participants effect). The increase with age appeared linear: the non-significant interaction between T1Age and ΔAge indicated that the change from Time 1 to Time 2 was similar for the whole range of ages at Time 1 (8–17 years). The marginally significant three-way interaction provided some evidence that participants of different ages at Time 1

Table 3
Parametric 95% confidence intervals for explanatory variables of the log odds of scoring above each level on items of the recursive thinking test for three models.

Explanatory variable	Model 1CI		Model 2CI		Model 3CI	
	LL	UL	LL	UL	LL	UL
One-loop	**−0.92**	**−0.46**	**−0.86**	**−0.31**	**−2.80**	**−0.56**
Two-loop	**−2.82**	**−2.27**	**−2.64**	**−2.09**	**−5.33**	**−2.87**
Male	−0.51	0.12	−0.93	0.09	−1.01	0.03
T1Age	**0.11**	**0.25**	−0.26	0.18	−0.32	0.14
ΔAge	**0.18**	**0.50**	**0.15**	**0.66**	**0.08**	**0.63**
T1Age × ΔAge	−0.09	0.05	−0.17	0.24	−0.16	0.25
One-loop × Male	−0.15	0.48	−0.07	1.02	**0.07**	**1.12**
Two-loop × Male	**0.02**	**0.77**	**0.14**	**1.29**	**0.21**	**1.36**
One-loop × T1 Age	−0.08	0.07	−0.13	0.35	−0.18	0.30
Two-loop × T1 Age	−0.11	0.07	−0.17	0.36	−0.25	0.27
One-loop × ΔAge	−0.25	0.08	−0.43	.09	**−0.59**	**−0.02**
Two-loop × ΔAge	−0.24	0.12	−0.36	0.20	−0.56	0.05
One-loop × T1 Age × ΔAge	−0.04	0.11	−0.28	0.12	−0.29	0.13
Two-loop × T1 Age × ΔAge	−0.00	0.16	−0.22	0.22	−0.21	0.24
DAT vocabulary					−0.01	0.06
DAT analogies					−0.05	0.02
One-loop × DAT vocabulary					−0.05	0.03
Two-loop × DAT vocabulary					−0.03	0.05
One-loop × DAT analogies					**0.03**	**0.10**
Two-loop × DAT analogies					**0.02**	**0.09**

Note: Model 1 was tested on the data from all participants. Model 2 and Model 3 were tested on the data from participants of 12 years and older at Time 1. T1 Age = age at Time 1 centered at the age of the youngest participant, ΔAge = the difference between the age at the time of testing and the age at Time 1, LL = lower limit, UL = upper limit. Confidence intervals of significant explanatory variables are printed in bold.

improved on different types of items: the increase from Time 1 to Time 2 in the log odds of scoring above each level became smaller for action items and larger for two-loop items as participants were older at Time 1.

Additional analyses were done to investigate the role of verbal abilities in the development of recursive thinking. In the second analysis (see Table 3, Model 2), scores on each of the 14 items at Time 1 and Time 2 were included for the 130 participants who had completed the DAT at both time points. The effect of type of item, the interaction between type of item and gender and the within-participants effect of age remained significant in this subsample. However, the effect of age at Time 1 (between-participants) was not significant in participants of 12 years and older.

The third analysis (see Table 3, Model 3) was done on data from the same participants and controlled for verbal abilities. The effect of type of item remained significant. The interaction between type of item and gender was now significant for both one-loop and two-loop recursion items. The difference in log odds between action items on the one hand and one-loop and two-loop recursion items on the other was smaller in males than in females. Moreover, there was a significant interaction between type of item and the score on DAT Analogies: higher scores on DAT Analogies were associated with higher log odds of scoring above each level on one-loop and two-loop recursion items. The within-participants effect of age remained significant when verbal abilities were controlled for. However, controlling for verbal abilities produced a significant interaction with type of item, indicating that improvement from Time 1 to Time 2 was smaller for one-loop recursion items than for action items.

4. Discussion

Most studies on recursive thinking- and other aspects of ToM (see Miller, 2012) -have focused on developments in the pre-school and childhood years. The present study aimed to extend the literature by testing the hypothesis that recursive thinking continues to develop beyond childhood and investigating whether this development levels off in early adolescence, as suggested by the results of Müller and Overton (2010), or continues throughout adolescence, as may be expected from the development of brain regions related to social cognition (Blakemore & Choudhury, 2006; Burnett & Blakemore, 2009; Nelson et al., 2005).

4.1. Psychometric properties

We first examined the psychometric properties of our revised version of the recursive thinking test developed by Miller et al. (1970) to establish that the test was suitable for use with adolescents. Factor analysis indicated that the revised test can be considered to measure a single dimension of recursive thinking with distinct levels of difficulty. The results showed that thinking about action and thinking about thinking precede thinking about thinking about thinking. However, the thresholds of transition to a higher score were not consistently lower for action items than for one-loop recursion items. This seems to indicate that most participants in our sample of 8–18-year-olds had mastered both thinking about action and thinking about thinking (i.e., one-loop recursive thinking).

The revised recursive thinking test showed good reliability. When considering all available participants, discriminant validity with verbal intelligence was better at Time 1, for 12–17-year-olds, than at Time 2, for 14–19-year-olds. However, correlations across a six-year age range may be inflated, because recursive thinking and verbal abilities are both related to age. Good discriminant validity was demonstrated at both time points when age was controlled for, indicating that the revised recursive thinking test measures something other than verbal intelligence. In future research, the relation with advanced ToM tasks could be examined to further validate that the recursive thinking test measures a social-cognitive ability.

4.2. Levels of recursive thinking

Both the present study and the study by Müller and Overton (2010) provided partial support for the scalability of recursive thinking. However, the levels of recursive thinking identified were not the same. The Rasch analysis by Müller and Overton (2010) showed a difference between action items and one-loop recursion items, but no clear difference between one-loop and two-loop recursion items. The present study, in contrast, showed no difference between action items and one-loop recursion items, whereas it did show a difference between one-loop and two-loop recursion items. In the study by Müller and Overton (2010), the difference in performance on action items and one-loop recursion items was largest in fifth grade (ages 10–12). Although the total age range was comparable, the proportion of participants under thirteen was somewhat smaller in the present study (28.8% vs 39.7%). This may explain the first discrepant finding. The second discrepancy may be related to differences in the educational diversity of the samples. The sample of Müller and Overton (2010) was recruited at private schools and may have been less diverse than the present sample. In view of the relation with verbal intelligence, educational level may, to some extent, explain the relatively high performance on two-loop recursion items in the study by Müller and Overton (2010).

The results of Müller and Overton (2010) suggested that adolescents have to learn that thinking can be recursive, but that there is essentially no difference between applying the principle of recursion once or twice. The results of the present study indicated that two-loop recursive thinking is a distinct step. This suggests that the number of times recursion is applied, or the number of perspectives that have to be taken into account, is relevant. Parallel to the difference between thinking about someone's actions and thinking about someone's thoughts there seems to be a difference between "seeing the world through someone else's eyes" and "reading a mind through someone else's eyes". An adolescent who can do the former, cannot necessarily do the latter.

Taken together, the results of the two studies support the developmental sequence of steps proposed by Miller et al. (1970): from thinking about contiguous people to thinking about action between people, to thinking about thinking and to thinking about thinking about thinking. The results also illustrate that, within a certain age group, scalability may be obscured by ceiling (or floor) effects for developments that span many years. Future research may demonstrate full scalability within a single study by sampling a uniform distribution of participants over a wide age range.

4.3. Development of recursive thinking

The present study was the first to provide longitudinal evidence on the development of recursive thinking. The results clearly support the assumption of continued development of recursive thinking in late childhood and adolescence. In the total sample, performance on the recursive thinking test was shown to increase with age both between participants (i.e., cross-sectionally) and within participants (i.e., longitudinally). In the subgroup of participants of 12 years and older, the between-participants effect of age was no longer significant, suggesting that the development of recursive thinking is less strongly related to chronological age in adolescence than in childhood. Nevertheless, the older subgroup showed significant improvement in recursive thinking over time. There was no sign that older participants showed less improvement (i.e., the interaction between T1Age and ΔAge did not approach significance in any of the analyses). In sum, recursive thinking appears to follow a linear developmental pattern, which does not level off before 18 years of age.

The pattern of development of recursive thinking shown here fits nicely with the pattern of development observed in childhood (Eliot et al., 1979; Landry & Lyons-Ruth, 1980; Miller et al., 1970; Oppenheimer, 1986; Veith, 1980). Scores according to the binary scoring system (see Supplementary Table 1) can directly be compared with the results from previous studies. This comparison shows that the youngest participants in the current study performed similar to their age-peers in the literature. The percentage correct for action items was around 85% in 9–12 year-olds, as has previously been reported for 4th to 6th graders (Miller et al., 1970; Müller & Overton, 2010; Veith, 1980). The percentage correct on one-loop recursion items, ranging from 68% to 79% in our 9–12-year olds, was higher than the 46–57% observed by Miller et al. (1970) for 4th to 6th graders, but not exceptional. In other studies, the percentage correct for one-loop items was 62.5 for 8–9-year-olds (Eliot et al., 1979) and 75 for 6th graders (Müller & Overton, 2010). The percentage correct on two-loop recursion items for our 9 and 10-year-olds (34 and 28, respectively) was also higher than the 17% correct Miller et al. (1970) reported for 4th graders, but in line with the finding by Eliot et al. (1979) that 8–9-year-olds scored more than 25% correct. The percentage correct for our 11 and 12-year-olds was the same as for 5th graders (Miller et al., 1970; Veith, 1980) and 6th graders (Miller et al., 1970): around 36%.

Because relatively high performance on action items has been found in late childhood, one might expect most improvement on one-loop and two-loop recursion items during adolescence. Indeed, Müller and Overton (2010) only found improvement in performance on two-loop recursion items between grades six and eight (ages 11–14). In the total sample of the present study, performance on all three types of items improved with age. However, there was a trend for older participants to show more improvement from Time 1 to Time 2 on two-loop recursion items than on action items.

The finding that boys did relatively well on two-loop recursion items compared to action items was unanticipated. Experimenters and raters informally observed that some boys seemed eager to demonstrate their ability to correctly describe the difficult two-loop items, while they tended to make sloppy mistakes on the action items (e.g., using 'thinking' instead of 'talking'). The finding is difficult to interpret theoretically and needs to be replicated before any conclusions can be drawn.

4.4. The role of verbal abilities in the development of recursive thinking

In the analysis using the data from participants of 12 years and older, a significant improvement from Time 1 to Time 2 remained when verbal abilities were controlled for, but it was reduced for one-loop recursive thinking. (For two-loop recursive thinking, the regression weight was in the same direction, but not significant). These findings indicate that some processes tapped by the recursive thinking test develop independently of verbal abilities, but also that improvement in one-loop recursive thinking is partly explained by the development of verbal reasoning. Verbal reasoning, measured by the analogies test, implies the understanding of abstract relations (De Wit & Compaan, 2005). As might be expected, this understanding contributes more to recursive thinking than vocabulary. In addition, verbal reasoning may become more important or even necessary for the development of higher levels of recursive thinking (i.e., one-loop and two-loop recursion). This is tentatively suggested by the higher correlations between verbal abilities and recursive thinking at Time 2 than at Time 1 and by the finding that age differences between participants contributed to explaining performance in the total sample, but not in participants of 12 years and older, while verbal reasoning did. However, this hypothesis remains to be tested.

Previous studies using a recognition version of the recursive thinking test have argued that the verbal nature of the cartoon description task leads to underestimation of recursive thinking. On recognition tests, 8 and 9-year-olds reached levels of performance close to those of our oldest participants (Eliot et al., 1979; Oppenheimer, 1986). Hence, one might expect that improvement on the cartoon description task in adolescence would be explained entirely by the development of verbal abilities. However, the present study also found improvement independent of verbal abilities. This finding is in line with some other studies showing that the spontaneous use of more advanced social cognition increases during adolescence. Barenboim (1978) argued that children and adolescents increasingly view other people as thinking agents. In describing people they knew, half of the 10-year-olds did not refer to thoughts, while the other half referred to thoughts about concrete things. The frequency of such references increased between ages 10 and 12. An increase in descriptions involving the person's thinking about thoughts or ideas occurred between 14 and 16 years. This may suggest that thinking about thinking becomes more automatic in adolescence. A finding by Dumontheil, Apperly and Blakemore (2010) that adolescents continue to improve in using another person's visual perspective to interpret instructions, although visual perspective taking per se is mastered in early childhood, suggests that other aspects of ToM develop in a similar way.

4.5. Directions for future research

Further research is needed to explore the suggestion that recursive thinking and other aspects of ToM become more automatic in adolescence. Second, the relation between these aspects needs to be clarified. Continued development throughout adolescence has not only been demonstrated for recursive thinking (in the present study), but also for affective perspective taking (Choudhury, Blakemore, & Charman, 2006) and the online usage of ToM (Dumontheil et al., 2010). A within-participants comparison of the development of these processes may clarify whether they share a common mentalizing component. Third, the relation between brain development and the development of recursive thinking may be investigated with a recognition version of the recursive thinking test. The results may parallel those of Sommer et al. (2010), who found that 10–12 year-old children and adults were both able to reason with false beliefs, but engaged different brain regions.

Another important direction for future research would be to study how the continued development of recursive thinking influences other developments in adolescence. For example, the growing ability to formulate recursive thoughts may enhance social performance, enabling more effective self-presentation, communication (Miller, 2012), and persuasion (Hill & Palmquist, 1978). Adolescents experience a shift in social orientation from parents to peers, spending more time with peers and increasingly valuing peer opinions (e.g., Nelson et al., 2005; Spear, 2000). Simultaneous improvements in recursive thinking may enable them to adjust their behavior to the expectations and opinions of peers. Alternatively, the realization that others may be thinking about them may increase fear of negative evaluation in adolescents (Bokhorst, Westenberg, Oosterlaan, & Heyne, 2008) and may even contribute to a stress response in social-evaluative situations (Van den Bos, Van Duijvenvoorde, & Westenberg, submitted). The revised recursive thinking test may be a useful instrument for future studies investigating such relations.

Finally, the finding that performance on the recursive thinking test continues to improve throughout adolescence calls attention to the conceptualization of recursive thinking. Miller (2012, p. 5) noted that: "Theory of mind has always been primarily a normative topic, in the sense that it concerns basic developments (such as mastery of false belief) that eventually almost all children demonstrate". Recursive thinking was likewise introduced as a sequence of steps in normal development (Miller et al., 1970). Although individual differences in the rate of development were acknowledged, all typically developing children were expected to demonstrate second-order recursive thinking at some point. The finding that individual differences in recursive thinking are much less stable than individual differences in (verbal) intelligence—a concept that was created to capture differences between individuals—lends some support to this view. As in the ToM literature, however, recent studies have demonstrated individual differences in higher order recursive thinking in adults (Kinderman et al., 1998; Nettle & Liddle, 2008; Paal & Bereczkei, 2007; Rutherford, 2004; Stiller & Dunbar, 2007; Sylwester et al., 2012), which were meaningfully related to their social functioning. These findings highlight that the development of recursive thinking may be more open ended.

4.6. Limitations

Limitations of the present study include attrition and the use of a convenience sample. However, the inclusion of participants from all levels of the educational system contributed to the sample's representativeness for the adolescent population in the Netherlands and other (Western European) countries. Another limitation is that data on verbal abilities were unavailable for participants under 12 years of age, because the DAT (De Wit & Compaan, 2005) is not suitable for children in elementary school. As a result, discriminant validity and the effect of verbal abilities on the development of recursive thinking could only be assessed for older participants. Among them, participants who were better at verbal reasoning were also better at one-loop and two-loop recursive thinking, but verbal abilities only partially explained improvement on the recursive thinking test over time. It is at present unclear whether verbal abilities influence the development of recursive thinking at a younger age or only become influential when children begin to think about thinking (about thinking). This may be clarified in future studies using a different test of verbal abilities.

Acknowledgements

We thank Anne Miers, Laura Compier-de Block, Ellen Middag, Manja Koenders and Masha Takens for their assistance in data collection and Senem Fincan and Jiska Koster for scoring the recursive thinking test.

Appendix A. Items of the revised recursive thinking test

Action items

1 The boy is thinking that the girl is talking to the (male) teacher.
2 The boy is thinking that he is talking to the girl.
3 The boy is thinking that the girl is talking to him.

One-loop recursion items

4 The boy is thinking that the girl is thinking of the (male) teacher.
5 The boy is thinking that he is thinking of the girl.
6 The boy is thinking that the girl is thinking of him.
7 The boy is thinking that the girl is thinking of herself.
8 The boy is thinking that he is thinking of himself.

Two-loop recursion items

9 The boy is thinking that the girl is thinking of the (male) teacher thinking of the (female) teacher.
10 The boy is thinking that he is thinking of the girl thinking of herself.
11 The boy is thinking that the girl is thinking of him thinking of her.
12 The boy is thinking that he is thinking of himself thinking of himself.
13 The boy is thinking that the girl is thinking of him thinking of himself.
14 The boy is thinking that he is thinking that the girl is thinking of him.

Appendix B. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.cogdev.2015.11.002.

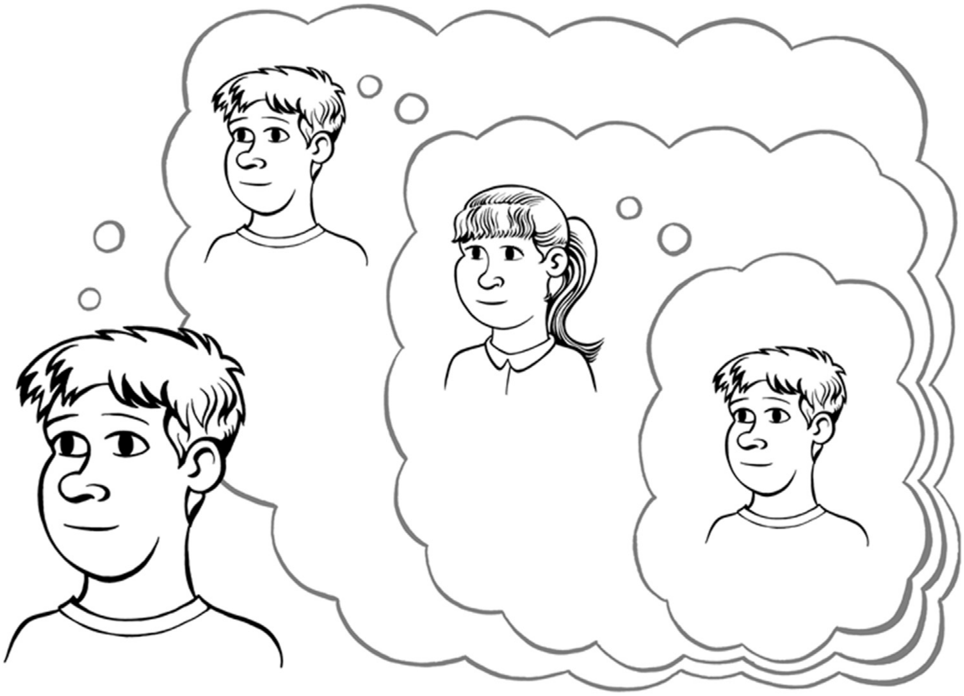

Fig. A1. Cartoon of item 14: The boy is thinking that he is thinking of the girl thinking of him.

References

Agresti, A. (2007). *An introduction to categorical data analysis*. Hoboken, NJ: John Wiley & Sons, Inc.

Barenboim, C. (1978). Development of recursive and nonrecursive thinking about persons. *Developmental Psychology*, *14*, 419–420. http://dx.doi.org/10.1037//0012-1649.14.4.419

Baron-Cohen, S., Wheelwright, S., Hill, J., Raste, Y., & Plumb, I. (2001). The reading the mind in the eyes test revised version: a study with normal adults, and adults with Asperger syndrome or high-functioning autism. *Journal of Child Psychology and Psychiatry*, *42*, 241–251. http://dx.doi.org/10.1017/S0021963001006643

Blakemore, S. J., & Choudhury, S. (2006). Development of the adolescent brain: implications for executive function and social cognition. *Journal of Child Psychology and Psychiatry*, *47*, 296–312. http://dx.doi.org/10.1111/j.1469-7610.2006.01611.x

Bokhorst, C. L., Westenberg, P. M., Oosterlaan, J., & Heyne, D. A. (2008). Changes in social fears across childhood and adolescence: age-related differences in the factor structure of the fear survey schedule for children-revised. *Journal of Anxiety Disorders*, *22*, 135–142. http://dx.doi.org/10.1016/j.janxdis.2007.01.014

Burnett, S., & Blakemore, S. J. (2009). The development of adolescent social cognition. *Annals of the New York Academy of Sciences*, *1167*, 51–56. http://dx.doi.org/10.1111/j.1749-6632.2009.04509.x

Choudhury, S., Blakemore, S. J., & Charman, T. (2006). Social cognitive development during adolescence. *Scan*, *1*, 165–174. http://dx.doi.org/10.1093/scan/nsl024

De Rooij, M. (2013). Standard regression models for repeated measurement and longitudinal data. *Unpublished manuscript*.

De Wit, J., & Compaan, E. (2005). *Differentiële Aanleg Test DAT NL Handleiding A Nederlandse bewerking*. London, UK: Harcourt Assessment.

Dumontheil, I., Apperly, I. A., & Blakemore, S. J. (2010). Online usage of theory of mind continues to develop in late adolescence. *Developmental Science*, *13*, 331–338. http://dx.doi.org/10.1111/j.1467-7687.2009.00888.x

Eliot, J., Lovell, K., Dayton, C. M., & McGrady, B. F. (1979). A further investigation of children's understanding of recursive thinking. *Journal of Experimental Child Psychology*, *28*, 149–157.

Emslie, H., Wilson, F. C., Burden, V., Nimmo-Smith, I., & Wilson, B. A. (2003). *Behavioural assessment of the dysexecutive syndrome for children (BADS-C)*. London, UK: Harcourt Assessment/The Psychological Corporation.

Ferguson, F. J., & Austin, E. J. (2010). Associations of trait and ability emotional intelligence with performance on theory of mind tasks in an adult sample. *Personality and Individual Differences*, *49*, 414–418. http://dx.doi.org/10.1016/j.paid.2010.04.009

Hedeker, D., & Gibbons, R. D. (2006). *Longitudinal data analysis*. New York, NY: Wiley.

Hill, J. P., & Palmquist, W. J. (1978). Social cognition and social relations in early adolescence. *International Journal of Behavioral Development*, *1*, 1–36. http://dx.doi.org/10.1177/016502547800100102

Jeon, I. H., Kim, K. R., Kim, H. H., Park, J. Y., Lee, M., Jo, H. H., et al. (2013). Attributional style in healthy persons: its association with 'Theory of Mind' skills. *Psychiatry Investigation*, *10*, 34–40. http://dx.doi.org/10.4306/pi.2013.10.1.34

Kinderman, P., Dunbar, R., & Bentall, R. P. (1998). Theory-of-mind deficits and causal attributions. *British Journal of Psychology*, *89*, 191–204. http://dx.doi.org/10.1111/j.2044-8295.1998.tb02680.x

Landry, M. O., & Lyons-Ruth, K. (1980). Recursive structure in cognitive perspective taking. *Child Development*, *51*, 386–394. http://dx.doi.org/10.2307/1129271

Miller, S. A. (2012). *Theory of mind: beyond the preschool years*. New York, NY: Psychology Press.

Miller, P. H., Kessel, F. S., & Flavell, J. H. (1970). Thinking about people thinking about people thinking about..: a study of social cognitive development. *Child Development*, *41*, 613–623. http://dx.doi.org/10.1111/j.1467-8624.1970.tb01018.x

Miers, A. C., Blöte, A. W., De Rooij, M., Bokhorst, C. L., & Westenberg, P. M. (2013). Trajectories of social anxiety during adolescence and relations with cognition, social competence, and temperament. *Journal of Abnormal Child Psychology*, *41*, 97–110. http://dx.doi.org/10.1007/s10802-012-9651-6

Müller, U., & Overton, W. F. (2010). Thinking about thinking–thinking about measurement: a Rasch analysis of recursive thinking. *Journal of Applied Measurement*, *11*, 78–90.
Nelson, E. E., Leibenluft, E., McClure, E. B., & Pine, D. S. (2005). The social re-orientation of adolescence: a neuroscience perspective on the process and its relation to psychopathology. *Psychological Medicine*, *35*, 163–174. http://dx.doi.org/10.1017/s0033291704003915
Nettle, D., & Liddle, B. (2008). Agreeableness is related to social-cognitive, but not social-perceptual, theory of mind. *European Journal of Personality*, *22*, 323–335. http://dx.doi.org/10.1002/per.672
Nippold, M. A. (1994). Third-order verbal analogical reasoning: a developmental study of children and adolescents. *Contemporary Educational Psychology*, *19*, 101–107. http://dx.doi.org/10.1006/ceps.1994.1009
Oppenheimer, L. (1986). Development of recursive thinking: procedural variations. *International Journal of Behavioral Development*, *9*, 401–411. http://dx.doi.org/10.1177/016502548600900309
Paal, T., & Bereczkei, T. (2012). Adult theory of mind, cooperation, Machiavellianism: the effect of mindreading on social relations. *Personality and Individual Differences*, *43*, 541–551. http://dx.doi.org/10.1016/j.paid.2006.12.021
Perner, J., & Wimmer, H. (1985). John thinks that Mary thinks that..: attribution of second-order beliefs by 5- to 10-year-old children. *Journal of Experimental Child Psychology*, *39*, 437–471. http://dx.doi.org/10.1016/0022-0965(85) 90051-7
Core Team, R. (2013). *R: a language and environment for statistical computing.* Vienna, Austria: R Foundation for Statistical Computing.
Rutherford, M. D. (2004). The effect of social role on theory of mind reasoning. *British Journal of Psychology*, *95*, 91–103. http://dx.doi.org/10.1348/000712604322779488
Saarni, C. (1979). Children's understanding of display rules for expressive behavior. *Developmental Psychology*, *15*, 424–429. http://dx.doi.org/10.1037//0012-1649.15.4.424
Sherman, M., & Le Cessie, S. (1997). A comparison between bootstrap methods and generalized estimating equations for correlated outcomes in generalized linear models. *Communications in Statistics-Simulation and Computation*, *26*, 901–925. http://dx.doi.org/10.1080/03610919708813417
Sim, J., & Wright, C. C. (2005). The kappa statistic in reliability studies: use, interpretation, and sample size requirements. *Physical Therapy*, *85*, 257–268.
Sommer, M., Meinhardt, J., Eichenmüller, K., Sodian, B., Döhnel, K., & Hajak, G. (2010). Modulation of the cortical false belief network during development. *Brain Research*, *1354*, 123–131. http://dx.doi.org/10.1016/j.brainres.2010.07.057
Spear, L. P. (2000). The adolescent brain and age-related behavioral manifestations. *Neuroscience and Biobehavioral Reviews*, *24*, 417–463. http://dx.doi.org/10.1016/s0149-7634(00) 00014-2
Stiller, J., & Dunbar, R. I. M. (2007). Perspective-taking and memory capacity predict social network size. *Social Networks*, *29*, 93–104. http://dx.doi.org/10.1016/j.socnet.2006.04.001
Sylwester, K., Lyons, M., Buchanan, C., Nettle, D., & Roberts, G. (2012). The role of theory of mind in assessing cooperative intentions. *Personality and Individual Differences*, *52*, 113–117. http://dx.doi.org/10.1016/j.paid.2011.09.005
Veith, D. L. (1980). Recursive thinking and the self-concepts of preadolescents. *Journal of Genetic Psychology*, *137*, 233–246.
Van den Bos, E., De Rooij, M., Miers, A. C., Bokhorst, C. L., & Westenberg, P. M. (2014). Adolescents' increasing stress response to social evaluation: pubertal effects on cortisol and alpha-amylase during public speaking. *Child Development*, *85*, 220–236. http://dx.doi.org/10.1111/cdev.12118
Van den Bos, E., Van Duijvenvoorde, A.C.K., & Westenberg, P. M. (submitted). Effects of adolescent socio-cognitive development on the cortisol response to social evaluation.
Westenberg, P. M., Bokhorst, C. L., Miers, A. C., Sumter, S. R., Kallen, V. L., et al. (2009). A prepared speech in front of a pre-recorded audience: subjective, physiological, and neuroendocrine responses to the Leiden Public Speaking Task. *Biological Psychology*, *82*, 116–124. http://dx.doi.org/10.1016/j.biopsycho.2009.06.005
Yu, C. Y., & Muthén, B. (2002). *Evaluation of model fit indices for latent variable models with categorical and continuous outcomes* Paper presented at the annual meeting of the american educational research association, New Orleans, LA.